Mobile and Wireless Communications Network Layer and Circuit Level Design

Edited by Salma Ait Fares and Fumiyuki Adachi

Mobile and Wireless Communications: Network layer and circuit level design

Edited by
Salma Ait Fares and Fumiyuki Adachi

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published 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.

We are IntechOpen,
the world's leading publisher of Open Access books
Built by scientists, for scientists

4,100+
Open access books available

116,000+
International authors and editors

120M+
Downloads

151
Countries delivered to

Our authors are among the
Top 1%
most cited scientists

12.2%
Contributors from top 500 universities

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Meet the editors

Salma Ait Fares received the B.S. degree from Mohammed V University, Rabat, Morocco, in 1999, and M.Sc. and Ph.D. degrees, respectively, in electrical engineering from University of Quebec in Trois Rivieres, in 2003, and in telecommunication from INRS-EMT, University of Quebec, Montreal, Qc, Canada, in 2007. She is currently JSPS postdoctoral dellow with Wireless Signal Processing & Networking Laboratory, Tohoku University, Sendai, Japan. Her research interests include cooperative relay network, OFDMA resource allocation, adaptive antenna arrays, sensor networks, and signal processing for telecommunication applications.

Fumiyuki Adachi received the B.S. and Dr. Eng. degrees in electrical engineering from Tohoku University, Sendai, Japan, in 1973 and 1984, respectively. In April 1973, he joined the Electrical Communications Laboratories of Nippon Telegraph & Telephone Corporation (now NTT) and conducted various researches on digital cellular mobile communications. From July 1992 to December 1999, he was with NTT Mobile Communications Network, Inc. (now NTT DoCoMo, Inc.), where he led a research group on Wideband CDMA for 3G systems. Since January 2000, he has been with Tohoku University, Sendai, Japan, where he is a Professor at the Dept. of Communications Engineering, Graduate School of Engineering. His research interest is in the area of wireless signal processing (multiaccess, equalization, antenna diversity, adaptive transmission, channel coding, etc.) and networking. He is an IEICE Fellow and an IEEE Fellow. He was a recipient of the IEEE Vehicular Technology Society Avant Garde Award 2000, IEICE Achievement Award 2002, Thomson Scientific Research Front Award 2004, Ericsson Telecommunications Award 2008, Telecom System Technology Award 2009, Prime Minister Invention Award 2010, British Royal Academy of Engineering Distinguished Visiting Fellowship 2011, KDDI Foundation Excellent Research Award 2012, VTS Conference Chair Award 2014, and C&C Prize 2014. He is listed in Highly Cited Researchers (http://highlycited.com/isi-highlycited.htm#table).

Preface

Mobile and wireless communications applications have clear impact on improving the humanity wellbeing. From cell phones to wireless internet to home and office devices, most of the applications are converted from wired into wireless communication. Smart and advanced wireless communication environments represent the future technology and evolutionary development step in home, hospitals, industrial, and vehicular and transportation systems. A very appealing research area in these environments has been the wireless ad hoc, sensor and mesh networks. These networks rely on ultra low powered processing nodes that sense surrounding environment temperature, pressure, humidity, motion or chemical hazardous, etc. Moreover, the radio frequency (RF) transceiver nodes of such networks require the design of transmitter and receiver equipped with high performance building blocks including antenna, power and low noise amplifiers, mixers and voltage controlled oscillators. Several challenges are facing nowadays researchers to design such building blocks while complying with ultra low power consumption, small area and high performance constraints. CMOS technology represents an excellent candidate to facilitate the integration of the whole transceiver on a single chip. However several challenges have to be tackled while designing using nanoscale CMOS technologies and require innovative idea from researchers and circuits designers. While major researcher and applications have been focusing on RF wireless communication, optical wireless communication based system has started to draw some attention from researchers for a terrestrial system as well as for aerial and satellite terminals. This renewed interested in optical wireless communications is driven by several advantages such as no licensing requirements policy, no RF radiation hazards, and no need to dig up roads besides its large bandwidth and low power consumption.

This second part of the book, Mobile and Wireless Communications: Key Technologies and Future Applications, covers the recent development in ad hoc and sensor networks, the implementation of state of the art of wireless transceivers building blocks and recent development on optical wireless communication systems. We hope that this book will be useful for the students, researchers and practitioners in their research studies.

This part consists of eighteen chapters classified in four corresponding sections.

1. *Network Aspects and Applications of Ad Hoc, Sensor and Mesh Networks*

2. *Antenna Design.*

3. *Wireless Transceivers Building Blocks in CMOS Technology.*

4. *Optical Wireless Communications.*

The first section contains five chapters related to Network Aspects and Applications of Ad Hoc, Sensor and Mesh Networks. In this section, the network layer design in cellular, ad hoc, sensor and mesh networks for specific applications have been presented.

The second section contains five chapters related to Antenna Design. In this section, different kind of UWB and microstrip antennas has been reviewed and developed. Their advantages, disadvantages, design technique, structure and application have been also covered.

The third section contains six chapters related to Wireless Transceivers Building Blocks in CMOS Technology. The focus of the contributions in this section, are the propose of a tunable polyphase filter structure, the development of wireless transceiver-on-a-chip on CMOS technology and the conception and development of several RFICs, such as, LNAs (Low Noise Amplifiers), mixer, and VCOs (Voltage Controlled Oscillators) in different applications.

The forth section contains two chapters related to Optical Wireless Communications. In this section, terrestrial free-space optical communication system has been addressed, in addition, a non-mechanical compact laser communications terminal for future applications has been proposed.

Section 1: *Network Aspects and Applications of Ad Hoc, Sensor and Mesh Networks*

Chapter 1 investigates the importance of CAC (Call Admission Control) in wireless networks for providing QoS guarantees. The key idea of this chapter, apart from offering a comprehensive study of CAC process in wireless networks, is to lay emphasis on the CAC method as a powerful tool to provide the desired QoS level to mobile users along with the maximization of network resource exploitation.

Chapter 2 describes the strategies developed so far to handle the problem of communication in strip-like topologies. Four approaches are presented in order to describe how each topology can be investigated. The first two are related to the network layer of ISO/OSI protocol stack, the third one proposes use of devices with directional antennas while the fourth one designs a MAC protocol based on synchronous transmit-receive patterns.

Chapter 3 introduces architecture for an all-to-all ad-hoc wireless network that satisfies the QoS requirements as well as power saving aspects. The power control algorithm which uses received signal strength measurements is also introduced.

Chapter 4 describes the wireless communication platform IQRF based on IQMESH protocol in terms of its advantages, strengths, limitations and specific implementations.

Chapter 5 reviews the automotive environment spread communication technologies and their areas of application, from short range to long range communication over several kilometers away.

Section 2: *Antenna Design*

Chapter 6 investigates passive wireless devices in the frequency range from almost DC to tens of Megahertz. This chapter provides a brief introduction to this technology, performance estimations in terms of powering range with respect to permitted signal levels and human exposure issues and analysis of the impact of conductive/dielectric materials in the vicinity of the passive wireless devices.

Chapter 7 introduces the UWB technology in terms of its history, definition, advantages and applications. An overview on UWB antennas including UWB planar monopole antennas and UWB printed antennas is presented. Two novel designs of UWB printed antennas are introduced and investigated in details where the structural properties and performance characteristics of these antennas are investigated.

Chapter 8 develops a micromachined aperture coupled patch antenna devices using polymer micromachining and micro-assembly methods to improve significantly the efficiency, gain and bandwidth of the devices over conventional microstrip patch antennas. The new fabrication method provides an alternative low cost packaging approach as compared to conventional LTCC and PCB technology.

Chapter 9 reviews different kind of microstrip antenna design mobile wireless communication systems such as microstrip antennas, microstrip array, compact and multiband microstrip antennas, broad band and UWB antennas, reconfigurable microstrip antennas and smart microstrip antennas. Their advantages, disadvantages, design technique, structure and application have been also covered.

Chapter 10 develops and demonstrates a large-signal model for GaN HEMTs, which accurately predicts trapping and self-heating-induced current dispersion and IMD. Detailed procedures for both small-signal and large-signal model parameter extraction has been presented.

Section 3: *Wireless Transceivers Building Blocks in CMOS Technology*

Chapter 11 proposes a tunable polyphase filter structure, which can be applied to synthesize multi-standard application filters. This tuning characteristic can be also used to compensate for the bandwidth drift due to mismatches.

Chapter 12 demonstrates the feasibility of low noise sensitivity 2.4GHz PLL for use in wireless communications in low cost LR-WPAN applications. The circuits have been fully integrated and implemented in 130nm CMOS technology. The proposed topology allows to realize much lower gain if it is required with a very simple calibration method.

Chapter 13 discusses enabling technologies for multi-gigabit spectrally efficient wireless communication systems in the E-band. The performance of state-of-the-art E-band wireless communication for high-capacity wireless networks has been evaluated. The analysis has been supported by experimental results on the prototypes.

Chapter 14 discusses the development of a 60-GHz wireless transceiver-on-a-chip on a 130-nm CMOS technology. The challenges and solutions for the design of 60-GHz components on CMOS including radio-frequency (RF) bandpass filter (BPF), power amplifier (PA), low-noise amplifier (LNA), mixers, voltage control oscillator (VCO) are described. These components are utilized to build the world's first all-integrated 60GHz wireless transceiver on CMOS which is also presented in this chapter.

Chapter 15 provides a guide to the RF building blocks of smart communication receivers in accordance with the present state of the art. The conception and development of several RFICs, such as, LNAs (Low Noise Amplifiers), mixer, and VCOs (Voltage Controlled Oscillators) in different applications have been introduced. The presented circuits can supply the necessities for many mobile applications, in particular, for SMILE (Spatial MultIplexing of Local Elements) front-end receiver circuitry.

Chapter 16 provides the fundamental background knowledge concerned with linear power amplifier design for high spectrum-efficiency wireless communications. In addition, the design considerations of the state-of-the art linear power amplifiers together with the design techniques operating at the gigahertz bands in CMOS technologies have been also covered.

Section 4: *Optical Wireless Communications*

Chapter 17 discusses the terrestrial FSO (Free-space optical) communication system from its basics to error performance based on OOK, PPM and SIM modulation schemes. The properties of the atmospheric channel have also been highlighted in terms of signal attenuation and scintillation.

Chapter 18 proposes a non-mechanical compact laser communications terminal for future applications. A laser beam is transmitted by selecting the laser pixel related to the direction of the optical signal received from the counter terminal. The beams are not deflected by a mechanical mirror. Instead, they are turned on and off one after the other in accordance with the direction from which optical signals are received.

Editors

Salma Ait Fares
Graduate School of Engineering
Department of Electrical and Communication Engineering
Tohoku University, Sendai, Japan
Email: aitfares@mobile.ecei.tohoku.ac.jp

Fumiyuki Adachi
Graduate School of Engineering
Department of Electrical and Communication Engineering
Tohoku University, Sendai, Japan
Email: adachi@ecei.tohoku.ac.jp

Contents

Call Admission Control in Mobile and Wireless Networks

Georgios I. Tsiropoulos, Dimitrios G. Stratogiannis
and Eirini Eleni Tsiropoulou
National Technical University of Athens
Greece

1. Introduction

The increasing demand for advanced multimedia services combined with the resource constraints of the wireless networks indicate the need of efficient admission control schemes to achieve a competent resource management combined with adequate Quality of Service (QoS) levels for end users. QoS provision in wireless networks is closely related to the exploitation of available network resources and the maximization of the number of users. Call Admission Control (CAC) is one of the key issues in wireless mobile communications, concentrating great interest in research work about QoS. CAC algorithms are employed to ensure that the admission of a new call into a resource limited network does not violate the Service Level Agreements (SLAs) concerning ongoing calls.

CAC schemes for wireless networks have been widely studied under different network architectures and network administrator policies. The objectives of the chapter are to present thoroughly the main concepts of CAC design and QoS provision in wireless and mobile networks. The study will focus on system and traffic analysis employed to model the complexity of communication traffic. In next generation networks where multiple Service Classes (SCs) with different QoS characteristics are supported, the various call types are classified into SCs with precise characteristics and QoS demands. Each SC call is treated differently depending on the criteria set according to the operating principles adopted for the admission procedure. CAC schemes handle multiple call stream flows corresponding to different priority levels providing an efficient mechanism to deal with different QoS necessities. The demanding environment of wireless communications poses numerous challenges in CAC design concerning the resource constraints, the connection quality, QoS requirements, SC prioritization, mobility characteristics and revenue optimization. Another critical issue in admission control is the performance evaluation, through appropriate metrics of the proposed schemes to assess the provided QoS. The metric studied most is Call Blocking Probability (CBP).

Finally, the last section of the chapter provides a broad classification of different design approaches and strategies considered for efficient admission control. CAC schemes are classified upon different rationales, used to apply call admission policy, aiding to an elucidatory synopsis of CAC under different network parameters. The majority of CAC

schemes base their admission criteria on an efficient resource management, accounted for either in terms of channels or bandwidth units. The methods proposed usually set thresholds related to the desirable QoS for high priority SCs and handoff calls. Other CAC schemes examine Signal to Noise Ratio (SNR) levels to determine an admission criterion satisfying the QoS demands of end users. Such schemes have to deal with propagation and mobility issues. Under high traffic network conditions, an efficiency enhancing module may be incorporated into the CAC schemes employed, renegotiating the resource allocation of ongoing calls. Through QoS re-negotiation and resource re-allocation, available resources can be retrieved and managed dynamically to serve a high priority SC call request.

2. Call Admission Control in Mobile and Wireless Communication

2.1 Call Admission Control (Definition and Operating Principle)

During the last decades the wireless communication networks users have been rapidly increased along with their demand for new multimedia services. The need for high speed communications is in contrast to the scarce spectrum resources allocated for wireless systems in international organizations. Therefore, a proficient radio resource management (RRM) is vital, to allot the existing network resources among contenting users, taking into consideration their needs and respective priorities as to provide them with the required QoS. More in detail, RRM functionality intents to improve system performance by maximizing the overall system capacity in the wireless network preserving at the same time the QoS characteristics of mobile users.

A crucial RRM mechanism essential for QoS provision applied on wireless networks is CAC. The key idea of Admission Control (AC) is to ensure the QoS of individual connections by appropriately managing the network resources. The main characteristics that an efficient AC policy should provide are the following: a) establish a robust priority assigning mechanism for handoff calls and calls of different SCs, b) exhibit a low CBP, c) allocate resources fairly, d) achieve a high network throughput and e) avoid congestion. Moreover, a proficient CAC scheme should avoid congestion and system outages due to overloading. The admission of a new call, according to the CAC scheme employed, should not violate the SLAs of ongoing calls. Admission decision is based on not only the available network resources but also the QoS requirements of the requesting and ongoing users. Hence, the decision should be taken considering multiple parameters such as the network characteristics, the service type, user mobility and the network conditions. In the case that the decision is positive, an appropriate quantity of network resources should be reserved to maintain the QoS of the new user. Thus, CAC is strictly related to resource allocation, channel and base station assignment, power control and resource reservation.

CAC problem can be considered as a multi-objective optimization problem that is maximizing the efficiency, utility and revenue of the network while at the same time complying with the users QoS requirements. The latter are provided by the users SLAs agreements. The admission criteria employed in the decision making part of the CAC scheme could be the Signal-to Interference Ratio (SIR), the ratio of bit energy to interference density ratio (E_b/I_0), the Bit Error Rate (BER), the Call Dropping Probability (CDP), the QoS at connection level as determined by the data rate and the delay bound. For instance, a CAC scheme may minimize the CBP by admitting a large number of call requests, provided that the BER violation probability does not exceed a satisfactory level ε_1 (Wu, 2005).

$$\Pr\{BER > BER_{thr}\} \le \varepsilon_1,$$

where BER_{thr} denotes the BER threshold. In the case that CBP values are available the above constraint can be rewritten as

$$CBP \le \varepsilon_2.$$

Generally, CAC schemes can be designed to provide different priority levels which correspond to the various SCs supported by the network.

2.2 Necessity for Call Admission Control and Quality of Service Provision

CAC algorithms are employed to ensure that the admission of a new call into a resource constrained network does not violate the SLAs of ongoing users. To decide whether to admit a new call or not, many factors are taken into consideration, most of them contradictory such as optimizing the use of radio resources, maximizing revenue, providing fairness, etc. Thus, CAC constitutes a mechanism which is used to determine the number of call connections so that different priorities are given among users with different QoS characteristics, network utilization is increased and congestion is prevented. Thus, when a call request from a mobile user is initiated, it may be accepted or blocked. The blocking probability, defined as the probability that a new call request is denied service by the network is called CBP and is subjected to the relevant decision made by the CAC scheme employed. Efficient CAC policies should achieve low CBPs.

Implementing practical CAC schemes is difficult; because traffic in communication networks is inherently chaotic and bursty, and traffic bursts are extremely difficult to be predicted. CAC schemes in wireless networks are complicated due to variable link quality and to users mobility. In particular, a call admitted in a certain cell may have to be handed off to a neighboring cell due to the users mobility. The main consideration in handoff procedures is to preserve the continuity of the call while at the same time offering at least the minimum acceptable QoS. During a call, a mobile user may cross several cell boundaries, thus requiring a corresponding number of successful handoffs. With regard to the handoff process, the new cell may not have any available resources to serve a handoff call, resulting in handoff failure commonly known as call dropping. In the literature, the probability that an ongoing call is terminated (dropped) is called CDP. It is widely accepted that users are more annoyed by call dropping than by call blocking; thus, efficient CAC schemes should keep CDP as low as possible. A simple way implemented in most CAC schemes, to achieve low CDP levels, is to assign higher priorities to handoff calls compared to new calls. Therefore, the admission criteria for new and handoff calls are different.

With regard to the number of active connections preserved, handoff schemes can be classified into hard handoff and soft handoff schemes. In the hard handoff schemes, a mobile terminal releases the channel from the original cell before its connection to the new Base Station (BS) is accomplished. Thus, a mobile terminal is connected to one BS at a time. In this case, the call is short-interrupted during the process of changing BS. In hard handoff schemes two ways leading to a handoff failure exist. The first is related to the way the handoff is implemented since if the old radio link is released before the network completes the assignment of a new channel, the call is dropped. This demonstrates the susceptibility of hard handoff schemes to the link transfer time. The second way may be attributed to the resource allocation mechanism since, if there are no channels available in the new cell, then the handoff call is forced-terminated.

In soft handoff schemes, the handoff process is triggered at the boundaries between neighboring cells. As cells in wireless systems overlap to assure complete coverage, the boundary areas may be served by more than one BS. Thus during the handoff, a mobile terminal may communicate with multiple BSs simultaneously, employing different radio links to achieve the communication with the network. When a channel from a BS is successfully assigned to a mobile terminal according to the specific handoff scheme QoS parameters, its originally occupied channels are released. In this case, the handoff procedure is insensitive to the duration of the handoff process, resulting in lower CDP compared to hard handoff schemes.

CAC schemes operate in real-time; hence, the algorithm used should be executed very fast. Moreover, the exact situation concerning the available resources at the BSs controller should be known as input data to the CAC algorithm. The design and implementation of a CAC scheme should be done very carefully aiming at minimizing false rejections and false admissions. A false rejection occurs when a call is rejected though the network has enough resources to serve it. In this case, optimization of network resources is not achieved, capacity is wasted and the operator's revenue is not maximized. On the other hand, a false admission occurs when a call request is accepted even if there are no available resources. In this case, the QoS level is not guaranteed and the CDP is increased, resulting in degradation of users satisfaction.

2.3 Challenges in Call Admission Control Design

The basic operation of CAC schemes is to decide whether a call should be admitted by the network or not. This decision is based on several criteria which are related to the network parameters and to the specific QoS characteristics of the call request. Although the QoS characteristics of the call are a priori determined, the network parameters are variable and adjustable in time. Thus, the CAC scheme employed should assure that the QoS characteristics of ongoing calls will not be violated throughout their whole duration. The factors employed in CAC schemes are presented below:

- Network load/resources: The limited network resources constitute a critical factor in CAC design. CAC schemes based on this criterion must know the resources available in each cell before the decision is taken. In this case, the network load after the admission of a new call must be predicted; if the predicted network load remains below a certain threshold, the new is admitted; otherwise, it is blocked. As handoff calls are treated differently by most CAC schemes, a set of channels may be reserved at each cell for handoff calls. Therefore, the admission of a new call is more difficult, as the respective threshold employed in most CAC schemes is lower, than the relevant threshold of handoff calls. These CAC schemes are widely known as Guard Channel (GC) schemes.
- Connection/link quality: Link quality is an essential parameter that should be taken into account when designing CAC in interference-limited wireless networks. Link quality refers to the radio link between the user terminal and the BS. For its estimation, the signal strength received at a mobile terminal and the interference caused to this link by other mobile terminals in the area are used. Thus, CAC schemes admit a new call if they can maintain the link quality of the admitted calls above a certain threshold. Otherwise, if the admission of a new call will result in an unacceptable deterioration of the link quality, the call is rejected. CAC schemes based on link quality usually employ the SIR

or the Signal-to-Noise-plus-Interference Ratio (SNIR) as an admission criterion; hence, they are called SIR or SNIR-based schemes.

- QoS requirements/call context: Since users may request services characterized by different QoS requirement with regard to mean throughput, mean delay, BER and bandwidth demands, the call requests are classified into various SCs. For every SC call request different admission criteria can be employed taking into consideration the respective QoS constraints and the network resources available. Thus, CAC schemes can be classified with regard to the number of the SCs supported. CAC scheme for single SC constituted a simple and appropriate model for first and second generation (2G) wireless networks, as they were mainly destined for voice services. The growing need for new services combined with the diffusion of new technologies, such as the 2.5 and 3G networks and also the Next Generation Networks (NGN), indicated the need to support multiple SCs with multimedia traffic and enhanced QoS characteristics. Thus, during the last decade, advanced CAC schemes supporting multiple SCs were introduced, classifying stream flows and call requests into different SC types according to their QoS characteristics. CAC design for multiple SCs is more challenging since different CAC criteria are employed for the SCs supported often resulting in high complexity and difficulties considering their implementation in practice.

- Call priority/SC prioritization: This CAC criterion is solely related to SC prioritization. Assigning higher priority to some SCs over the rest is a common technique in CAC schemes for multiple SC networks. In particular, it is widely accepted that Real Time (RT) services have higher priority over Non-Real Time (NRT) ones, e.g. a voice call is considered of higher priority compared to internet browsing. Moreover, different priorities can be assigned even within the same SC reflecting the differentiation among different user classes, stemming from subscription fee policy. Also, higher priorities are assigned to handoff calls or to calls related to emergency services. Different priority levels reflect different CAC criteria, which are more strict for low priority SC calls and relaxed for high priority ones.

  Prioritization schemes can be implemented mainly through: channel borrowing, queuing and reservation schemes. In channel borrowing schemes, if a cell has all its channels reserved, it can borrow channels from neighboring cells to serve high priority SC calls. In queuing schemes, if a cell has all its resources occupied, a high priority call request is set into a queue until resources, sufficient to accommodate the call request are released in the cell,. Queuing schemes can be applied either to high priority call requests or to all incoming call requests (regardless of their priority). In the latter case their position into the queue is adjusted according to the respective requests priority. On the other hand, the reservation schemes were first used to give priority to handoff calls by permanently reserving on a permanent basis a number of channels exclusively for serving handoff requests. These schemes have been extended to support multiple SCs by assigning different priority levels through reserving channels for high priority SC calls.

- User's mobility characteristics: Users mobility is a critical factor in wireless networks as users travel across multiple cells; thus, the traffic in the cells is variable and it cannot be precisely predicted as an active terminal may move from one cell to a neighboring one, resulting in calls handoff. If a handoff call cannot be served by the BS of the new cell, it is dropped increasing the call dropping rate. Since users are more sensitive to call dropping than to call blocking, CAC schemes are employed to reduce the handoff failure

probability. Most schemes in the literature assign higher priorities to handoff calls resulting in less strict admission conditions for the admission of handoff calls. These schemes are the same with the prioritization schemes mentioned above with the difference that they are destined to prioritize handoff calls.

- Transmission rate: CAC schemes are employed to guarantee the minimum bandwidth requirements for ongoing calls. Moreover, every SC call may also have a maximum bandwidth requirement. Based on the available resources, a CAC scheme aims at providing the highest possible bandwidth between the minimum and maximum requirement to every call and, at the same time, reducing CBP. To this end, certain CAC schemes incorporate QoS renegotiation, a mechanism which is activated when the cell resources of network cell are not sufficient, to reduce the transmission rate of ongoing calls, as much as required for the admission of a new call. The reduced transmission rate may be increased when resources are released due to the termination of a call.

- Revenue optimization: By applying a proper network utilization policy, an efficient CAC scheme may provide a high revenue for the network operator. On the other hand, there are strict limitations imposed by the total bandwidth constraints and the QoS guarantee through the SLAs. Any admitted call contributes to the revenue increase but it may also cause a penalty if the QoS of ongoing calls is deteriorated. The reward may be represented by the number of users or the portion of occupied bandwidth whereas the various penalties may be defined via the probability of QoS deterioration. To determine in real time the optimum equilibrium between reward and penalties is a rather complicated problem. The relevant CAC schemes are named revenue optimization or economic CAC schemes.

- Fairness in resource assignment: The main drawback of CAC schemes basing their admission criterion on the call priority is that high priority calls often monopolize the network resources. This results in a severe blocking of low priority calls and, consequently, in high CBP levels for the low priority traffic flows. This is observed not only in networks supporting multiple SCs where different priority levels are assigned to each SC, but also among different users in the same SC with different SLAs and mobility characteristics. Specific CAC schemes exist which take into consideration fairness criteria based on various network parameters, such as the network throughput or the CBP achieved, to ensure that no SC or user class dominates the network resources.

3. Mobile & Wireless Networks Modeling and Traffic Analysis

3.1 Traffic Model and System Analysis

The majority of the studies concerning CAC in wireless networks make certain standard assumptions to provide a tractable analysis. Most system models were obtained through common traffic theory and have been extended to cellular networks. These networks are not necessarily represented by these traffic models, since users mobility and the emerging multimedia services necessitate new teletraffic assumptions and models that take into account the new aspects of wireless networks.

A fundamental assumption in modeling wireless networks with regard to CAC is that the new call arrivals in a cell follow the Poisson distribution, that is, the new calls arrive in cell i according to a Poisson distribution with rate $\lambda_{n,i}$. If the network is assumed homogeneous, the arrival rate is the same for every cell and the analysis may be limited to only a single cell.

The handoff call arrivals in cell i are also assumed to follow the Poisson distribution with rate $\lambda_{h,i}$. Such an assumption is not so obvious in the case of handoff calls as the handoff traffic is solely related to the user mobility characteristics. It has been proven (Chlebus & Ludwin, 1995) that this assumption is valid provided there is no blocking in the network. As this is an ideal case, in the same work a blocking scenario is assumed to examine how accurate is the assumption that handoff arrivals follow the Poisson distribution. The results indicate that through this approximation of the real situation the performance exhibited is satisfactory. Moreover, in the same work it is argued without providing the proof that in blocking environment handoff traffic is a smooth process which means that the variance is less than the mean value. It must be noted that in Poisson distribution the variance is equal to the mean value. Apart from adopting the Poisson distribution for modeling the arrival rate, other traffic models have been proposed in the literature. In (Rajaratnam & Takawira, 2000) the authors suggest that the call arrival process in wireless networks should be modeled according to general distribution (Rajaratmam & Takawira, 1999). Moreover, they have shown that the handoff traffic is a smooth process if the channel holding times follow the exponential distribution.

The channel holding time is defined as the time that a channel is assigned to a call in a certain cell. The channel is released after the call is either terminated or handed off to a neighboring cell. Another important term in wireless networks is the call holding time (also referred to the literature as service time or Requested Call Connection Time, RCCT), defined as the total connection time originally requested by a call. The call holding time varies according to the type of the call, as calls belonging to different SCs may also have different durations. How long a call stays in a cell is another fundamental parameter in wireless networks and is widely called as Cell Residence Time (CRT) or cell dwell time. CRT is mainly dependent on users mobility characteristics and on the geometry of the cells.
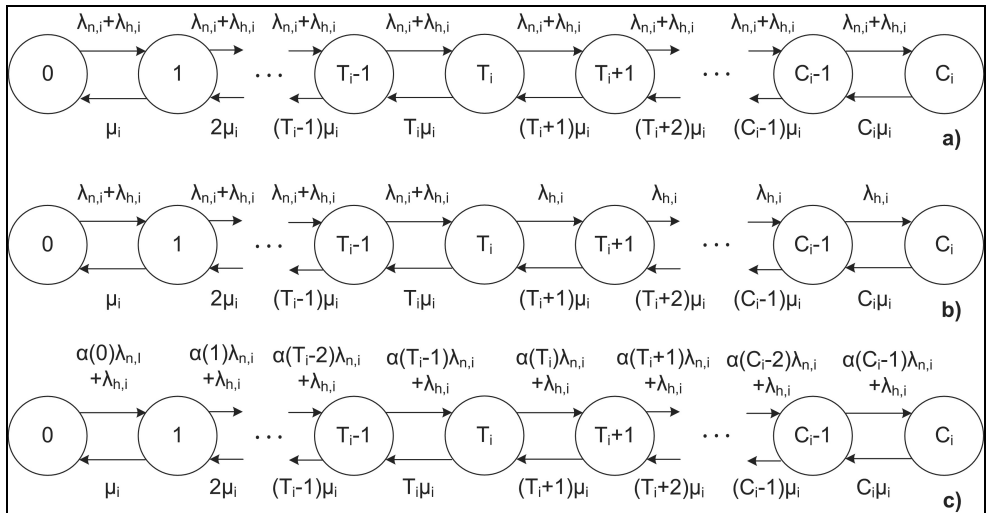

Fig. 1. Transition diagrams considering network state. a) Complete resource sharing scheme, b) Guard Channel scheme and c) Fractional Guard Channel scheme.

The majority of the analyses existing in the literature assume that the channel holding times follow the exponential distribution for both new and handoff calls. However, the channel holding time follow the exponential distribution, only under certain conditions investigated in (Fang, Chlamtac, & Lin, 1998), where it is proven that channel holding time follows the exponential distribution if the CRT is also exponentially distributed. In all the other cases, the channel holding time cannot be modeled according to the exponential distribution whereas neither the handoff traffic nor the new incoming traffic flow follow the Poisson distribution. Some researchers adopt other distributions to model the channel holding time such as the lognormal (Jedrzycki & Leung, 1996) and general distribution (Rajaratmam & Takawira, 1999). Although modeling the cell residence time and the channel holding time is not straightforward, most researchers model both these characteristics through the exponential since under this assumption the relevant analysis becomes tractable yielding analytical formulas for the CBP. A more rigorous approach is beyond the scope of this chapter; therefore, both new and handoff incoming traffic will be assumed as Poisson arrivals whereas the channel holding time and cell dwell time in cell i will be modeled through the exponential distribution with mean $1/\mu_i$.

In a complete resource sharing scheme (Lai, Misic, & Chanson, 1998) a call is admitted as long as there are sufficient network resources to accommodate the call; otherwise it is rejected. The same policy is applied for new and handoff calls. By defining the state of a cell i at time t $\{c_i(t)\,|\,t{\geq}0\}$ as the number of occupied channels in cell, the cell state can be modeled as a Continuous-Time Markov Chain (CTMC). If the respective number of channels is C_i, the system model is a typical $M/M/C_i$ queue (Figure 1a). Note that to adopt the $M/M/C_i$ model an assumption should be made that when the network operates under congestion a new or handoff call arrival is blocked. This assumption reduces the analysis from $M/M/C_i/K$, where K is maximum number of calls waiting to be served, into $M/M/C_i$, where no buffer is used. The truncated state space of cell i is represented by S_i, where

$$S_i=\{n_i;\ 0{\leq}n_i{\leq}C_i\}.$$

Let $\pi(n_i,n_i')$ be the transition rate from state n_i to state n_i', where $n_i,\ n_i' \in S_i$. Then, the transition probabilities for adjacent states are obtained from

$$\pi(n_i,n_i+1)=\lambda_{n,i}+\lambda_{h,i}$$
$$\pi(n_i,n_i-1)=n_i\mu_i.$$

Based on the transition diagram depicted in Figure 1a the following global balance equation is derived

$$(\lambda_{n,i}+\lambda_{h,i})\ p(n_i)=(n_i+1)\mu_i\ p(n_i+1),$$

where $p(n_i)=\lim_{t\to\infty}\mathrm{Prob}[c_i(t)=n_i]$ denotes the steady state probability that the number of ongoing calls in cell i is n_i, $n_i=0,1,\ldots,C_i$. From the global balance equation the steady state probabilities are obtained from

$$p\left(n_i\right) = p\left(0\right)\rho^{n_i}\big/n_i!\,,\quad 0{\leq}n_i{\leq}C_i,$$

where $\rho=(\lambda_{n,i}+\lambda_{h,i})/\mu_i$ is the traffic intensity and p(0) is the normalization factor defined as

$$p(0) = \left[\sum_{n_i=0}^{C_i} \frac{\rho^{n_i}}{n_i} \right]^{-1} .$$

A new call destined for cell i is blocked if all its channels are occupied; hence, the new call blocking probability in cell i is given by

$$P_n{}^b(i)=p(C_i).$$

Since no prioritization for handoff calls has been assumed in this general analysis, the handoff failure probability $P_h{}^b(i)$ in cell i should be equal to $P_n{}^b(i)$. Therefore

$$P_h{}^b(i)= P_n{}^b(i)= p(C_i).$$

This analysis may be extended to multiple SCs and multiple cells but it proves very complicated (Li & Chao, 2007) as the transition diagram has multiple dimensions rendering the corresponding global balance equation difficult to solve. In most cases, different traffic flows, each of which corresponds to a specific SC, are considered to be independent; therefore, multiple one-dimensional transition diagrams are obtained, reducing the complexity of the problem. An interesting and mathematically robust analysis concerning this problem is provided in (Li & Chao, 2007) where expressions for CBPs, handoff rates and QoS (also called grade of service) are obtained in closed form.

As previously mentioned, in the case of multiple independent SCs, the previous analysis is carried out separately for every SC. Consider U SCs with arrival rates

$$\lambda_{u,i}(n_{u,i})=\lambda_{nu,i}(n_{u,i})+\lambda_{hu,i}(n_{u,i})$$

and death rates

$$\mu_{u,i}(n_{u,i})=\mu_{u,i}n_{u,i} \ ,$$

where u=1,…,U and $\lambda_{nu,i}(n_{u,i})$, $\lambda_{hu,i}(n_{u,i})$ are the respective call arrival rates for new and handoff u SC calls in cell i and $\mu_{u,i}$ is the respective mean cell residence time. The steady state probability of having $n_{u,i}$ channels in cell i occupied by u SC calls is

$$p_u\left(n_{u,i}\right) = \left(\frac{\lambda_{nu,i} + \lambda_{hu,i}}{\mu_{u,i}}\right)^{n_{u,i}} \frac{1}{n_{u,i}!} p_u(0) ,$$

where $p_u(0)$ is the normalization factor given by

$$p_u(0) = \left[\sum_{n_{u,i}=0}^{C_i} \left(\frac{\lambda_{nu,i} + \lambda_{hu,i}}{\mu_{u,i}} \right)^{n_{u,i}} \frac{1}{n_{u,i}!} \right]^{-1}.$$

Considering now the total number of SCs supported in cell i, the truncated state space is

$$S'_i = \{n_i = (n_{1,i}, n_{2,i}, \ldots, n_{U,i}); n_{1,i} + n_{2,i} + \ldots + n_{U,i} \leq C_i\}.$$

The steady state probability that the network is at state n_i is given by

$$\mathbf{p}_{u,i}(n_i) = \mathbf{p}_{u,i}(0) \prod_{u=1}^{U} \left(\frac{\lambda_{nu,i} + \lambda_{hu,i}}{\mu_{u,i}} \right)^{n_{u,i}} \frac{1}{n_{u,i}!},$$

where $\mathbf{p}_{u,i}(0)$ is the normalization factor given by

$$\mathbf{p}^{-1}_{u,i}(0) = \sum_{n_i \in S_i} \prod_{u=1}^{U} \left(\frac{\lambda_{nu,i} + \lambda_{hu,i}}{\mu_{u,i}} \right)^{n_{u,i}} \frac{1}{n_{u,i}!}.$$

For the complete resource sharing scheme, the CBP and CDP for u SC in cell i are equal to the probability that cell i is under congestion, e.g. all its channels are occupied. Thus, the corresponding probability is given by

$$\mathbf{P}_h^b(u,i) = \mathbf{P}_n^b(u,i) = \mathbf{p}(n_i^*),$$

where $n_i^* = \left(n_{1,i}^*, n_{2,i}^*, \ldots, n_{U,i}^* \right); n_{1,i}^* + n_{2,i}^* + \ldots + n_{U,i}^* = C_i$.

In literature, there are two approaches concerning the whole network problem where J cells are supported, i=0,1,…,J. In the first case, the network may be assumed as homogeneous; then, it suffices to examine one cell only with its results representing the whole network behavior. Therefore, the CBP and CDP determined previously for cell i apply for the whole network. In the second case, the network traffic is not uniformly distributed over all the cells supported; then, appropriate analysis should be carried out to determine the admission failure probabilities. This analysis is analytically presented in (Li & Chao, 2007) where additional QoS network parameters are examined.

3.2 Service Classes Classification

Former generations of wireless networks used simple traffic shaping schemes where all traffic was shaped uniformly by rate. This model was realistic as only one service (voice calls) was offered. As modern wireless networks offer a variety of services, the incoming traffic should be classified into different traffic types. Each traffic type is called SC and the procedure followed to determine in which class a new call request falls into is called classification. Each SC has its own QoS characteristics with regard e.g. to bitrate, packet delay, duration etc. Therefore, each SC should be treated differently to differentiate the

service destined for the user. Despite the increased complexity due to multiple SCs supported by the network, the control mechanisms are more flexible in resource allocation management and QoS provision. Apart from different QoS characteristics for each SC concerning physical and network layer, different priority levels are applied to different SCs supported employing certain policies. This SC prioritization is usually based on the QoS requirements, the pricing policy followed by the administrator and the users SLAs. This differentiation of the incoming calls can be utilized by a network operator to treat the various SC calls in different ways with regard to bandwidth allocation, call admission process, pricing policy, etc.

A usual classification is the differentiation of the incoming calls into two general SCs, real-time SCs and non-real-time SCs (Tsiropoulos, Stratogiannis, Kanellopoulos, & Cottis, 2008). This classification is primarily based on the latency characteristics of the various calls. In general, there is a deadline for a data packet to be delivered to its destination. If for a certain call this requirement is strict or lenient; the call is characterized as RT call or NRT, respectively. In modern wireless networks supporting multimedia traffic a broader call classification is required. Apart from taking into consideration the latency of each call, additional QoS requirements are considered such as the bandwidth required and the call duration. Therefore, calls are classified into multiple SCs (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008) such as voice, messaging, internet browsing and file transfer, teleconference etc.

Recent trends in traffic control classify the incoming calls into three SCs: Premium, Gold and Silver (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008) (Guo & Chaskar, 2002). Premium SC calls are assigned with the highest priority level and they are offered the negotiated bandwidth all the time, regardless of congestion, interference or degradation of channel quality. A lower priority level is assigned to Gold SC calls and the lowest one to Silver SC calls. The resources are allocated to calls according to the respective SC priority level. Thus, in case of congestion, Premium SC calls are still served under their initially requested QoS characteristics, whereas Gold and Silver SC calls are subject to QoS degradation in proportion to their priority levels so that congestion is mitigated. According to this classification of calls, each mobile user may associate each application with either of three SCs according to its QoS expectations and the pricing scheme applied (Guo & Chaskar, 2002). Thus, for a certain call, say a voice call, a user may associate it with the Premium SC, whereas other users may associate a voice call with the Gold SC. Regardless of the call classification scheme adopted, call classification simplifies the network analysis and enhances QoS provision, as calls are managed in groups and not independently

3.3 Efficiency and Performance Evaluation

CBP Estimation

The common criteria employed to evaluate the performance of all the CAC schemes proposed are CBP and CDP. When the assumptions made allow the application of Markov chain analysis, analytical formulas for CBP and CDP are derived (Li & Chao, 2007; Fang & Zhang, 2002; Tsiropoulos, Stratogiannis, Kanellopoulos, & Cottis, 2008). Therefore, the assessment of the CAC schemes employed can be based on these criteria. In measurement-based CAC schemes, CBP and CDP are estimated by measuring the calls blocked or dropped, respectively, during a predefined time window. The CAC scheme proposed in the

literature aim at reducing as much as possible both these probabilities by adopting an appropriate decision making procedure. Moreover, the QoS requirements of the ongoing calls should be satisfied at the same time providing prioritization to handoff calls. Both CBP and CDP are mainly dependent on the input traffic load, the number of ongoing calls, the bandwidth requirements of each call and the policy applied for handoff calls (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008).

In single SC networks the assessment of CAC schemes with regard to their failure probabilities is focused on handoff prioritization (Fang & Zhang, 2002; Yavuz & Leung, 2006). The divergence between CBP and CDP becomes greater as the policy for handoff prioritization becomes stricter. In GC schemes this is realized by lowering the threshold level T whereas in fractional schemes the probability $\alpha(n_i)$ becomes lower. To measure the prioritization achieved between new and handoff calls an appropriate priority index (PRIN) is defined as the fraction of CBP to CDP

$$\mathrm{PRIN} = \frac{\mathrm{CBP}}{\mathrm{CDP}}.$$

To achieve handoff prioritization, PRIN should be higher than unity, as the CBP should be greater than CDP.

A similar analysis is applied in multiple SC networks. Apart from prioritizing handoff calls, different SCs should also be assigned with different priority levels. Thus, considering that a SCs should have priority over SC u+1, (where u,u+1∈U), then CBP_u and CDP_u should be lower than CBP_{u+1} and CDP_{u+1}. Therefore, the divergence between failure probabilities among different SCs is more critical in multiple SC networks. The PRIN index can be modified to incorporate the prioritization level of different SCs. In particular,

$$\mathrm{PRIN}\left(\mathrm{u,u'}\right) = \frac{\mathrm{CBP_u} + \mathrm{CDP_u}}{\mathrm{CBP_{u'}} + \mathrm{CDP_{u'}}},$$

measures the prioritization achieved among u and u' SC, u,u'∈U, where $\mathrm{PRIN(u,u')} > 1$ if $\mathrm{u < u'}$ or $\mathrm{PRIN(u,u')} < 1$ if $\mathrm{u > u'}$.

4. Call Admission Control Design Approaches

4.1 Classification of Call Admission Control Schemes

CAC schemes can be classified into general categories based either on the criteria considered in the decision part of the CAC scheme or on specific design characteristics. The admission criteria considered by CAC schemes are usually related to various QoS parameters and have been discussed earlier. Each design characteristics has its own advantages and disadvantages. The selection among different CAC approaches should be based upon the wireless technology used, the SCs supported and the geographical characteristics of the region where the network is installed.

With regard to the centralization level of CAC schemes, they are classified into centralized, distributed or collaborative. In centralized schemes, CAC is implemented at the Mobile Switching Center (MSC) which is responsible for handling the services supported by the

network. The information from the BS of a cell must be aggregated at the MSC where the admission decision is taken; then, the BS is commanded to act accordingly. The main advantage of centralized CAC schemes is their high efficiency, but the high level of complexity along with the increased redundancy due to the control data required, makes them unrealistic in practice. In distributed CAC schemes the decision making part is installed at the BS of each cell and completes the CAC procedure independently of the other cells. Therefore, they are more reliable and more easily implemented. However, they are less efficient as they lack global information about the network parameters, information available only in centralized CAC schemes. The collaborative schemes (O'Callaghan, Gawley, Barry, & McGrath, 2004), constitute a promising hybrid design option. In such schemes, information concerning resource allocation and admission control is exchanged between neighboring cells, though the decision is taken by the BS of each cell. Hence, the advantages of centralized and distributed CAC schemes are combined in effective powerful architecture offering high efficiency and increased reliability. The main disadvantage of collaborative schemes is the high overhead required.

CAC schemes can also be discriminated into traffic-descriptor-based - also called proactive - or measurement - based - also called reactive. In the former scheme, the admission decision is based on the traffic pattern which is available for the application of these schemes, which check whether the already reserved bandwidth increased by the bandwidth demand of the new call exceeds the cell capacity. In this case, the call is blocked otherwise it is admitted. The most common traffic-descriptor-based CAC scheme is the simple sum scheme (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008) (Jamin, Shenker, & Danzig, 1997) which simply ensures that the sum of the requested resources does not exceed the cell capacity. A new call with maximum bandwidth demand r_α is admitted under the condition that the already occupied bandwidth demand increased by r_α remains below the cell capacity μ, that is if:

$$v + r_\alpha \leq \mu.$$

As multimedia traffic is bursty in nature, traffic-descriptor-based CAC schemes overestimate the bandwidth demands since traffic descriptors specify the maximum bandwidth demand in each call which is rarely used. On the other hand, traffic-descriptor-based CAC schemes are very simple; ergo they are widely used by switch and router vendors.

In measurement-based CAC schemes the decision making module employs the actual network characteristics such as the actual traffic load, the packet error rate etc which are appropriately measured and, consequently, realistic. Some interesting measurement-based CAC schemes considered in (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008; Jamin, Shenker, & Danzig, 1997) are based on the actual traffic flow, the occupied bandwidth, the network load and packet loss accompanied with revenue award. The fundamental parameter in measurement-based CAC schemes is the measuring mechanism itself, in other words how the parameter employed in the CAC procedure is measured (Jamin, Shenker, & Danzig, 1997; Warfield, Chan, Konheim, & Guillaume, 1994; Dziong, Juda, & Mason, 1997; Casetti, Kurose, & Towsley, 1996). The measurement procedure is performed either by directly measuring the proper network parameter every sampling period following a time-window policy (Jamin, Danzig, Shenker, & Zhang, 1997), or by computing a relevant average value based on current and/or previous measurements (Jamin, Shenker, & Danzig,

1997; Floyd, 1996). Most CAC schemes employed in CDMA systems are designed according the measurement-based technique (Stasiak, Wisniewski, & Zwierzykowski, 2005).

Another interesting classification of CAC schemes can be made based on the amount of information available at the decision making module. This information may include the number of available or occupied cell channels, the total bandwidth allocated to ongoing users, the mean packet delay for each traffic flow etc. If this information can span over the whole network, the scheme is characterized as global. As expected, these schemes achieve high efficiency but exhibit exceptional complexity and require the exchange of a huge amount of information among the network cells. If the information exchange is done within a limited area including at least the neighboring cells of the cell under consideration, the CAC scheme is called semi-local. These schemes achieve also high efficiency and are less complex compared to global ones but they still require a lot of information exchange. Apart from information exchanging schemes, local CAC schemes exist which base their admission decision only on the information concerning a specific cell. Local schemes are simple to implement; however, they are less efficient compared to global or semi-local schemes since they do not take into account that, due to users mobility the load of a cell is influenced by the load of the neighboring cells.

Many CAC schemes are available in literature proven to achieve the optimal solution to the CAC problem, according to the inputs for the admission decision process. However, optimal CAC schemes often require a high computational power for their implementation, due to the large number of states associated with the Markov Decision Problem (MDP). The large scale of the problem and the multiple interdependent network parameters employed in optimal CAC schemes result in high complexity and increased processing time. Thus, theoretically optimal CAC schemes are not applicable in practice, as the admission decision must be taken instantaneously upon a call request. As an alternative approach to optimal CAC schemes, suboptimal CAC schemes have been proposed which operate online with significantly lower complexity. Suboptimal CAC schemes obtain a near-optimal solution to the CAC problem, usually by employing intellectual techniques (heuristic functions, alternative approaches, etc) to reduce the complexity of the original problem.

CAC schemes can be also classified based on information granularity (Jain & Knightly, 1999) which depends on the traffic model adopted, the spatial distribution of network users and the way network information is obtained. CAC schemes may adopt a specific users mobility pattern, otherwise a simple resource policy for mobile users will be used. In the first case, the exact knowledge of the users mobility characteristics, such as direction and velocity, helps to predict the handoff traffic load destined to each cell. The spatial users distribution may be uniform or non-uniform; consequently, the wireless network is considered as homogeneous or non-homogeneous respectively. Information can be obtained at each cell for either each call or each SC stream flow. As the information about the network increases, the complexity of the CAC scheme increases along with its efficiency.

An additional classification of CAC schemes can be done based on the differentiation of the data rates between the uplink and the downlink. Unlike traditional voice services, the demand for bandwidth between uplink and downlink is asymmetric in many multimedia applications. In relevant systems, if the CAC scheme employed allocates equal bandwidth to both uplink and downlink traffic, system capacity might be limited by the downlink traffic (Yang, Feng, & Kheong, 2006); then resources are used inefficiently, bandwidth is wasted and efficiency performance of the CAC scheme is low. Some CAC schemes adopt a joint

admission policy, by accepting a new call provided that enough resources can be allocated to both uplink and downlink according to the QoS characteristics of the new call.

4.2 Call Admission Control based on Signal Quality

In modern wireless access technologies, interference poses critical constraints concerning mainly the signal quality. This situation has an impact not only on network conditions but also on systems capacity. Particularly, in CDMA wireless networks interference is the dominant factor affecting their performance in terms of capacity and QoS provision to end users. Thus, the SINR is an adequate metric of the signal quality. CDMA-based air interfaces are mainly influenced by interference caused by other users from the same network instead of Gaussian noise, so the noise effect is usually neglected focusing mainly on SIR.

Therefore, CAC schemes implemented for interference - limited networks employ as admission criterion either the interference levels caused by a new incoming call or the signal quality levels achieved. Hence, interference based CAC schemes admit new calls only if the SNR/SIR values can maintain a minimum signal quality level. The SNR/SIR levels correspond to predefined QoS levels for new and ongoing users. This simple approach offers a tool to reduce interference in wireless networks, while on the other hand is constitutes an efficient admission criterion.

Two simple SIR-based solutions were first proposed by (Liu & Zarki, 1994) for controlling the signal quality. This is achieved by checking the achievable SIR value by the new call. The call is admitted provided that this value is higher than the minimum SIR value. Both implemented schemes are based on the residual capacity of the cell formulating each time an appropriate admission criterion. In the first scheme the residual capacity of the network is defined as

$$R_k = \left[\frac{1}{SIR_{th}} - \frac{1}{SIR_k} \right],$$

where SIR_k is the uplink SIR value in a cell k and SIR_{th} is the threshold value that imposes whether a call is admitted or not. The residual capacity of the cell is calculated when a new user arrives and if is greater than zero the incoming call is admitted otherwise the call is rejected. The second proposed algorithm follows the same rationale taking also into account the impact of admitting one call on cell k itself and its adjacent cells C(k) as well. This is done by encompassing an interference coupling parameter β in the above definition of the residual capacity leading to

$$R_{k,j} = \left[\frac{1}{\beta} \left(\frac{1}{SIR_{th}} - \frac{1}{SIR_j} \right) \right], \quad j \in C(k).$$

These simple algorithms were evolved taking into account inter-cell interference. In residual capacity estimation, the parameter $L_m(j,k)$ is used, representing the predicted additional intercell interference. The use of this parameter results into service enhancement in terms of reduced CBPs. QoS guarantees are also provided by using certain bounds for threshold,

maintaining specified levels for blocking rate (Kim, Shin, & Lee, 2000). The residual capacity is estimated by

$$R_{k,j} = \left[\frac{1}{SIR_{th}} - \frac{1}{SIR_j} - \frac{1}{L_m(j,k)} \right] j \in C(k).$$

The SINR is a sufficient metric for the signal quality providing also supplementary information. In particular, through SINR measurement BER can be estimated given the coding and modulation techniques applied. Additionally, from SNR/SIR values E_b/N_0 and energy per bit to interference density ratio (E_b/I_0) can be calculated respectively. CAC schemes control the signal quality employing as a decision criterion the equivalent constraints applied on E_b/N_0. Such SIR based schemes were developed for multicode CDMA systems (Ayyagari & Ephremides, 1998). The proposed scheme orders users based on the E_b/N_0 required. Before admitting a user, the CAC scheme checks whether the user can be assigned a minimum number of codes (corresponding to the minimum transmission rate) beginning from the user with the lowest E_b/N_0. If the minimum number of codes can be assigned without violating the constraints on E_b/N_0, the user is admitted. In the next step, additional codes are attempted to be assigned to the user to increase the transmission rate up to the maximum designated rate. Then, the system proceeds to the next user until all users are checked. Therefore, every user is either admitted and allocated multiple codes or rejected due to system infeasibility.

The SIR-based CAC schemes offer a reliable platform for dynamic adaptive admission control facing the problems of non-stationary and non-uniform traffic. This is realized through either the adoption of a local adaptive scheme employing the traffic parameters and the channel characteristics or a global scheme collecting information from all the neighbouring cells. Dziong & Jia (1996) proposed a relative framework that estimates the mean and variance of the interference level measuring them via a Kalman filter and predicts mean and variance of the interference concerning the new call The total interference is assessed as the sum of the estimated interference of all calls, the predicted interference emerged by the admission new call and the reservation threshold (a reservation for the estimated interference variance and errors). Then, the admission decision is taken comparing the total interference to the maximum tolerable interference threshold I_{max}.

Apart from interference – limited CDMA networks CAC schemes based on signal quality control were developed for TDMA systems. This kind of system suffers from signal quality constraints especially when a tight frequency reuse plan is employed to increase the number of available channels. Thus, in TDMA networks SIR based CAC schemes are utilized to guarantee SIR_{min} since cochannel interferers are in close distance one each other, causing significant problems in signal quality. The CAC scheme checks all available time slots in the cell and admits the user if at least one time slot has SIR > SIR_{min}. If SIR remains below SIR_{min} throughout the call duration, the call is reassigned to a different time slot that satisfies the SIR requirements offering the desired QoS (Haleem, Avidor, & Valenzuela, 1998). Moreover, CAC schemes remain a prevailing solution for hybrid T/CDMA systems combining multicode techniques with SIR based CAC (Casoni, Immovilli, & Merani, 2002).

SIR based CAC schemes performance can be improved employing optimization techniques. In such schemes a constrained optimization problem is formulated employing an objective

function including the signal quality constraints. Through the optimization of the objective function, the system capacity is maximized avoiding high failure probabilities. An optimum CAC policy for multiservice networks is proposed by (Singh, Krishnamurthy, & Poor, 2002), accomplishing to minimize blocking probability for one class, while using as constraints blocking rates for the supported SCs and the SIR conditions. The problem of optimization is solved by means of linear programming. Applying these techniques the efficiency of the CAC scheme is increased leading to an advanced resource management

4.3 Handoff and Service Class Prioritization Schemes

Complete resource sharing schemes cannot guarantee a certain QoS level for handoff calls, especially during network congestion, as the total numbers of channels are allocated to new and handoff calls with no discriminations. Therefore, GC schemes, also called reservation schemes, have been first proposed by Hong and Rappaport (1986) in the mid 80s to prioritize handoff calls over new ones. The basic concept behind this scheme is to reserve a certain number of channels only for serving handoff calls. The rest of the channels are available for both new and handoff calls. Thus, assuming that the total number of channels in cell i is C_i and that the number of channels available for common use are T_i ($T_i < C_i$), the number of channels dedicated to handoff calls only is $C_i - T_i$. Therefore, T_i operates as a threshold, implying that a new call is admitted provided that at least $C_i - T_i + 1$ channels are available, whereas the admission of a handoff call requires only one available channel. Under the same network assumptions as with the single SC complete resource sharing scheme the transition diagram for the CTMC is shown in Figure 1b, represented by the same truncated state space S_i. The transition probabilities to adjacent states in GC schemes are given by

$$\Pi(n_i, n_{i+1}) = \lambda_{n,i} + \lambda_{h,i} \quad \text{for } 0 \le n_i \le T-1$$
$$\Pi(n_i, n_{i+1}) = \lambda_{h,i} \quad \text{for } T \le n_i \le C-1$$
$$\Pi(n_i, n_{i-1}) = n_i \mu_i.$$

Based on the global balance equations provided by the CTMC depicted in Figure 1b, the following expression is derived for the state space probability $p(n_i)$, that is

$$p(n_i) = p(0) \left(\frac{\lambda_{n,i} + \lambda_{h,i}}{\mu_i} \right)^{n_i} \frac{1}{n_i!}$$

$$p(n_i) = p(0) \left(\frac{\lambda_{n,i} + \lambda_{h,i}}{\mu_i} \right)^{T_i} \left(\frac{\lambda_{h,i}}{\mu_i} \right)^{n_i - T_i} \frac{1}{n_i!}$$

where p(0) is the normalization factor defined as

$$p(0) = \left[\sum_{n_i=0}^{T_i-1} \left(\frac{\lambda_{n,i} + \lambda_{h,i}}{\mu_i} \right)^{n_i} \frac{1}{n_i!} + \sum_{n_i=T_i}^{C_i} \left(\frac{\lambda_{n,i} + \lambda_{h,i}}{\mu_i} \right)^{T_i} \left(\frac{\lambda_{h,i}}{\mu_i} \right)^{n_i - T_i} \frac{1}{n_i!} \right]^{-1}.$$

According to the CG CAC scheme, a new call is blocked if the total occupied channels in cell i exceeds T_i-1; hence, the CBP is given by

$$P_n{}^b(i)= p(T_i)+ p(T_{i+1})+\ldots+ p(C_i).$$

As GC schemes provide prioritization to handoff calls, a handoff call is dropped if all channels of cell i are occupied; consequently CDP is given by

$$P_h{}^b(i)= p(C_i).$$

If T is equal to C, the CG scheme reduces to a complete resource sharing scheme. The main issue in GC schemes is the evaluation of the threshold value T that minimizes CBP under the constraint set for CDP (Ramjee, Towsley, & Nagarajan, 1997) (Harine, Marie, Puigjaner, & Trivedi, 2001). The stricter the constraint for handoff calls, the better is the QoS provided by the network. On the other hand, lower values of CDP result in higher values for CBP. Hence, there is a tradeoff between CBP and CDP, which is reflected in the evaluation of the threshold level.

As modern wireless networks support multiples SCs with different QoS characteristics, sophisticated CAC schemes should be employed to guarantee the various QoS levels. Thus, GC schemes, which were originally destined to prioritize handoff calls over new ones, may be modified by adopting different thresholds for each SC calls (new and handoff) in the relevant decision process. These schemes are commonly known in the literature as multi-threshold bandwidth reservation schemes (Haung & Ho, 2002) (Chen, Yilmaz, & Yen, 2006). In such schemes, two kinds of prioritization can be applied, SC priority and call type priority. The former characterizes the type of the service and addresses the problem of prioritizing different SC calls by employing different thresholds for the calls of each SC. The analysis is similar to the one presented above for the complete resource sharing scheme supporting multiple SCs.

The different SCs supported are usually RT and NRT SCs. In other cases, the SCs supported may be more than two (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008) each with its own QoS requirements. The main problem of multiple threshold schemes is the exact evaluation of the threshold values. Several algorithms exist in the literature for the evaluation of thresholds mainly based on revenue optimization, but most of them are computationally expensive and difficult to solve (Chen, Yilmaz, & Yen, 2006). The evaluation of the proper threshold is a multi-constrained problem, where multiple network limitations should be taken into account.

The fixed number of channels reserved for handoff calls or high priority SC calls by GC or multi-threshold schemes, respectively, is determined mainly based on a priori knowledge of the traffic patterns. As they are static in nature, they lack adaptability thereby being unable to adapt to the network variations. In realistic situations, the call arrival rates are not static and cannot be efficiently described through traffic patterns. Hence, a network with time-varying traffic requires dynamic CAC schemes which usually achieve a better performance. Dynamic (also called adaptive) GC schemes are based either on mobility prediction employing the users mobility, the direction and the location of the mobile terminal, or on measurements of certain network parameters such as CBP, CDP or the available resources. Their main objective is to adapt the thresholds employed by the CAC scheme to optimize a given objective function formulated through a penalty – revenue approach concerning the

admission or rejection of incoming calls, a network utilization optimization policy, etc. Both the prediction of mobility and the measurement-based dynamic CAC schemes are strongly dependent on the efficiency of the mobility estimation algorithm or the measurement technique employed, respectively. The extensive results presented in the literature indicate that dynamic GC schemes outperform static GC schemes (Yu & Leung, 1997) (Bartolini & Chlamtac, 2001). On the other hand, dynamic GC schemes are more complex compared to static schemes, resulting in increased exchange of information between the BSs and large processing load.

Another variation of the GC schemes presented above is the Fractional Guard Channel (FGC) schemes. The FGC scheme was first proposed by Ramjee (1996) and was proven to be more general than the GC schemes. The basic idea of the FGC scheme lies in the admission of an arriving new call in cell i under a certain probability $\alpha(n_i)$, $0 \leq \alpha(n_i) \leq 1$, where the parameter $n_i = 0, 1, \ldots, C_i$, depends on the number of busy channels in cell i. In general, this probability varies with the network status. If $\alpha(0) \geq \alpha(1) \geq \ldots \geq \alpha(C_i)$, the stream of new call becomes thinner as the number of occupied channels increases. Handoff calls are unhinderly admitted unless no channels are available. CBP and CDP vary significantly as the number of occupied channels varies. On the other hand, in FGC schemes the application of admission probabilities eliminates the use of threshold in admission decision providing a smoother variation of CBP and CDP.

Consider the operation a FGC scheme under the same network assumptions as in the CG schemes previous referred to. The network states, represented by the same truncated state space S_i, can be modeled according the CTMC depicted in Figure 1c. As observed in Figure 1c, the transition probabilities for adjacent states in FGC schemes are obtained from

$$\pi(n_i, n_{i+1}) = \alpha(n_i)\lambda_{n,i} + \lambda_{h,i}$$
$$\pi(n_i, n_{i-1}) = n_i\mu_i.$$

Consequently, the global balance equation is obtained applying the fundamental rule of the birth-death process in Figure 1c "rate up equals rate down" (Cooper, 1981) as follows

$$(\alpha(n_i)\lambda_{n,i} + \lambda_{h,i})\, p(n_i) = n_{i+1}\mu_i\, p(n_{i+1}).$$

Solving the global balance equation, the steady state distribution of the FGC scheme is obtained from

$$p(n_i) = \frac{\prod_{k=0}^{n_i} \alpha(k)\lambda_{n,i} + \lambda_{h,i}}{\mu_i^{n_i} n_i!} p(0) \quad \text{for } 0 \leq n_i \leq C_i,$$

where

$$p(0) = \left[\sum_{n_i=0}^{C_i} \frac{\prod_{k=0}^{n_i} \alpha(k)\lambda_{n,i} + \lambda_{h,i}}{\mu_i^{n_i} n_i!} \right]^{-1}.$$

CBP and CDP may be obtained from

$$P_n^b(i) = \sum_{n_i=0}^{C_i} \left(1 - \alpha(n_i)\right) p(n_i)$$

and

$$P_h^b(i) = p(C_i),$$

respectively.

When $\alpha(0)=\alpha(1)=\ldots=\alpha(T_i)=1$ and $\alpha(0)=\alpha(T_{i+1})=\ldots=\alpha(C_i)=0$, the FGC scheme reduces to the simple GC scheme with T_i acting as threshold for new calls. Also, setting all fractional probabilities equal to unity, e.g. $\alpha(0)=\alpha(1)=\ldots=\alpha(C_i)=1$, the complete resource sharing scheme is obtained. Thus, the FGC schemes are proven more general with the GC and the resource sharing schemes being special cases. Due to the incorporation of the fractional probability $\alpha(n_i)$, the FGC schemes may be extended to comply with network administrator specifications and SCs QoS requirements in multimedia wireless networks.

The FGC schemes are further employed to prioritize high priority SCs in multimedia wireless networks. These schemes are also known as thinning or probabilistic CAC schemes (Wang, Fan, & Pan, 2008) (Tsiropoulos, Stratogiannis, Kanellopoulos, & Cottis, 2008). Each SC call is assigned with its own probability. SCs of higher priority are assigned with higher probabilities. In this case, the previous analysis of FGC schemes can be generalized according to the analysis employed for multi-SCs in complete resource sharing and GC schemes (Wang, Fan, & Pan, 2008).

4.4 Bandwidth Adaptation and Quality of Service Renegotiation

Wireless networks support a variety of services which can be classified into rate-adaptive applications and constant bitrate (CBR) services. In such services, e.g. voice calls, a bandwidth increase beyond the standard requirement will not improve the respective QoS. On the other hand, in rate adaptive services users specify, at their connection request, the minimum and maximum bandwidth required. Apart from specifying the bandwidth range required by every SC, rate variations may originate from the dynamic nature of the wireless environment along with the mobility of user terminals. Thus, in modern wireless networks bandwidth adaptation algorithms are employed to improve network utilization and guarantee the QoS of ongoing calls, assigning the minimum bandwidth required. When the network conditions are favorable and enough resources are available, they may be assigned to ongoing rate-adaptive users according to two general strategies based on SCs priorities (Li & Chao, 2007). According to the first strategy, the available resources are fairly assigned to all ongoing users without taking into account any priorities. According to the second, resources are first assigned to SC calls of high priority; until the resources are exhausted or all high priority SC calls have taken the maximum bandwidth required. If resources are still available, the scheme assigns them to SC calls of the next high priority. The procedure continues until all resources are exhausted or all calls are served with their maximum bandwidth demand. Apart from taking into account priorities, resource assignment in rate-adaptive services may be performed through more complicated schemes. In (Sen, Jawanda, Basu, & Das, 1998), an optimal resource assignment strategy is proposed for maximizing the

total revenue obtained, while in (Sherif, Habib, Nagshineh, & Kermani, 2000), an adaptive resource allocation scheme is proposed to maximize bandwidth utilization and attempt to provide fairness with a generic algorithm.

In general, when a call arrives in a certain cell, the network may either have enough resources to provide bandwidth between the minimum and the maximum demand or be congested, that is, it cannot provide the minimum bandwidth requested by the new call. In the first case the call is admitted, whereas in the second, bandwidth adaptation CAC algorithms, also known as rate-adaptive schemes, are applied to determine an optimal resource allocation aiming at serving as many users as possible while reducing the admission failure probability. This is accomplished by reducing the rate of some users when possible as much as required to accommodate the new call. In some bandwidth adaptation CAC schemes, this procedure is followed only for handoff or for call requests of high priority SCs (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008; Lindemann, Lohmann, & Thümmler, 2004). However, it should be mentioned that user rates cannot be reduced below the minimum rate values required to assure QoS; thus, when all users operate at their lowest bandwidth requirement, a new call request will be rejected. Rate degradation may be enforced according to a prioritization or to a non-prioritization scheme. In the former, the rate degradation policy is first applied to the SC calls of the lowest priority. If the resources released are still not sufficient for the admission of a new call, the calls of the next priority level are examined. In the non-prioritization schemes, all calls served with higher rates than their minimum bandwidth demand reduce their rate to admit the call request. A useful metric in QoS renegotiation CAC schemes is the degradation ratio which is defined as the ratio of the number of degraded calls to the number of ongoing calls (Kwon, Choi, Bisdikian, & Naghshineh, 1999). Moreover, the degradation probability can be determined though network measurements. Higher or lower degradation probabilities correspond to how aggressive a CAC design approach is.

The reverse procedure is followed when enough available resources exist to offer higher rates to ongoing calls. This rate upgrade policy can be applied in two ways. According to the first one, a rate adaptive resource allocation scheme is employed to exploit the available resources (Li & Chao, 2007). According to the second one, the calls having had their rate decreased more recently are the first calls to have their rate restored (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008). If enough available resources still exist, a resource allocation scheme is employed to assign them to ongoing users.

QoS renegotiation, especially rate degradation must be used carefully and should be the last step of a CAC scheme in an effort to acquire the resources necessary for the admission of a new call. There are many applications, such as voice calls or video streaming, with rates that cannot be reduced (QoS degradation) at not noticeable levels by the user. A drawback of rate adaptive CAC schemes comes up when a network operates near congestion. Then, a certain number of calls may undergo multiple rate degradations followed by respective rate restorations, as call requests arrive and ongoing calls are terminated, respectively. As users are sensitive to rate fluctuations, it is preferable to employ appropriate thresholds in the rate upgrade procedure which implies that a rate upgrade is done only if the available resources remaining after the upgrade are above the threshold (Tragos, Tsiropoulos, Karetsos, & Kyriazakos, 2008).

5. Conclusion

In this chapter the importance of CAC in wireless networks for providing QoS guarantees has been investigated. CAC algorithms are important for wireless networks not only for providing the expected QoS requirements to mobile users, but also to maintain network consistency and prevent congestion. To address the problem of CAC the main term of QoS has been firstly examined. Different QoS levels supported by the network correspond to the various SCs offered to mobile users. Each SC has its own requirements and specifications which should be met to offer a satisfactory QoS to end users. Thus, various challenges arise in designing efficient CAC schemes that have been determined and thoroughly investigated in the present chapter. An important aspect of CAC schemes, to measure their appropriateness for a given network, is the criteria which should satisfy. The main idea of CAC scheme classification is that different schemes apply individual criterion on admission procedure. Moreover, various system architectures exist which demand different CAC schemes, properly designed to adapt to system characteristics. Furthermore, the concerns of the network administrator should be taken into account, applying the policy needed for revenue optimization and maximum resource exploitation through CAC. Analytical models for the most common CAC schemes have been exhibited. An efficient CAC scheme should achieve low failure probabilities, high network resources exploitation, fairness in resource allocation among different users and revenue optimization. To evaluate the performance of CAC schemes studied according to these aspects, various efficiency criteria have been presented. The key idea of this chapter, apart from offering a comprehensive study of CAC process in wireless networks, is to lay emphasis on the CAC method as a powerful tool to provide the desired QoS level to mobile users along with the maximization of network resource exploitation.

6. References

Ahmed, M. H. (2005). Call Admission Control in Wireless Networks: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials, 7* (1), 50-66.

Ahn, C. W., & Ramakrishna, R. S. (2004). QoS provisioning dynamic connection-admission control for multimedia wireless networks using a Hopfield neural network. *IEEE Transactions on Vehicular Technology*, *53* (1), 106-117.

Ayyagari, D., & Ephremides, A. (1998). Admission Control with Priorities: Approaches for Multi-rate Wireless Systems. *IEEE International Conference on Universal Personal Communications 1998 (ICUPC'98). 1*, pp. 301-305. Florence: IEEE.

Bartolini, N., & Chlamtac, I. (2001). Improving call admission control procedures by using hand-off rate information. *Wireless Communications and Mobile Computing*, *1* (3), 257-268.

Casetti, C., Kurose, J. F., & Towsley, D. F. (1996). A new algorithm for measurement-based admission control in integrated services packet networks. *Fifth International Workshop on Protocols for High-Speed Networks (PfHSN '96). 73*, pp. 13 - 28. Sophia Antipolis: Chapman & Hall, Ltd.

Casoni, M., Immovilli, G., & Merani, M. L. (2002). Admission control in T/CDMA systems supporting voice and dataapplications. *IEEE Transactions on Wireless Communications, 1* (3), 540-548.

Chen, I. R., Yilmaz, O., & Yen, I. L. (2006). Admission Control Algorithms for Revenue Optimization With QoS Guarantees in Mobile Wireless Networks. *Wireless Personal Communications: An International Journal*, *38* (3), 357-376.

Chlebus, E., & Ludwin, W. (1995). Is Handoff Traffic Really Poissonian? *Fourth IEEE International Conference on Universal Personal Communications* (pp. 348-353). Tokyo: IEEE.

Cocchi, R., Shenker, S., Estrin, D., & Zhang, L. (1993). Pricing in computer networks: motivation, formulation, and example. *IEEE/ACM Transactions on Networking*, *1* (6), 614-627.

Cooper, R. B. (1981). *Introduction to Queueing Theory.* New York: Elsevier North Holland Inc.

Dziong, Z., & Jia, M. (1996). Adaptive traffic admission for integrated services in CDMAwireless-access networks. *IEEE Journal on Selected Areas in Communications*, *14* (9), 1737-1747.

Dziong, Z., Juda, M., & Mason, L. G. (1997). A framework for bandwidth management in ATM networks-aggregate equivalent bandwidth estimation approach. *IEEE/ACM Transactions on Networking*, *5* (1), 134-147.

Evci, C., & Fino, B. (2001). Spectrum management, pricing, and efficiency control in broadbandwireless communications. *Proceedings of the IEEE*, *89* (1), 105-115.

Fang, Y., & Zhang, Y. (2002). Call admission control schemes and performance analysis in wirelessmobile networks. *IEEE Transactions on Vehicular Technology*, *51* (2), 371-382.

Fang, Y., Chlamtac, I., & Lin, Y. B. (1998). Channel Occupancy Times and Handoff Rate for Mobile Computing and PCS Networks. *IEEE Transactions on Computers*, *47* (6), 679-692.

Floyd, S. (1996). *Comments on measurement-based admission control for controlled-load services.* Berkeley: Lawrence Berkeley Laboratory.

Guo, Y., & Aazhong, B. (2000). Call Admission Control in Multi-class Traffic CDMA Cellular System Using Multiuser Antenna Array Reciever. *IEEE Vehicular Technology Conference (VTC '00). 1*, pp. 365-369. Japan: IEEE.

Guo, Y., & Chaskar, H. (2002). Class-based quality of service over air interfaces in 4G mobilenetworks. *IEEE Communications Magazine*, *40* (3), 132-137.

Haleem, M. A., Avidor, D., & Valenzuela, R. (1998). Fixed wireless access system with autonomous resource assignment. *The Ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications 1998 (PIMRC'98). 3*, pp. 1438-1442. Boston: IEEE.

Harine, G., Marie, R., Puigjaner, R., & Trivedi, K. (2001). Loss formulas and their application to optimization for cellular networks. *IEEE Transactions on Vehicular Technology*, *50* (3), 664-673.

Haung, Y. R., & Ho, J. M. (2002). Distributed call admission control for a heterogeneous PCS network. *IEEE Transactions on Computers*, *51* (12), 1400-1409.

Hong, D., & Rappaport, S. S. (1986). Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, *35* (3), 77-92.

Hou, J., Yang, J., & Papavassiliou, S. (2002). Intergration of Pricing with Call Admission Control to Meet QoS Requirements in Cellular Networks. *IEEE Transactions on Parallel and Distributed Systems*, *13* (9), 898-910.

Hwang, Y. H., & Noh, S. K. (2005). A call admission control scheme for heterogeneous service considering fairness in wireless networks. *Fourth Annual ACIS International Conference on Computer and Information Science 2005 (ICIS 2005)* (pp. 688-692). Jeju Island, South Korea: IEEE Computer Society.

Ibrahim, W., Chinneck, J. W., & Periyalwar, S. (2003). A QoS-based charging and resource allocation framework for next generation wireless networks. *Wireless Communications and Mobile Computing*, *3* (7), 895-906.

Jain, R. K., Chiu, D. M., & Hawe, W. R. (1984). *A quantitative measure of fairness and discrimination for resource allocation in shared computer systems.* Maynard, Massachusetts: Digital Equipment Corporation.

Jain, R., & Knightly, E. W. (1999). A framework for design and evaluation of admission controlalgorithms in multi-service mobile networks. *INFOCOM '99* (pp. 1027-1035). New York: IEEE.

Jamin, S., Danzig, P. B., Shenker, S. J., & Zhang, L. (1997). A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks. *IEEE/ACM Transactions on Networking*, *5* (1), 56-70.

Jamin, S., Shenker, S. J., & Danzig, P. B. (1997). Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service. *IEEE INFOCOM'97. 3*, pp. 973-980. Kobe: IEEE.

Jedrzycki, C., & Leung, V. C. (1996). Probability distribution of channel holding time in cellulartelephony systems. *IEEE 46th Vehicular Technology Conference 1996 (VTC '96). 1*, pp. 247 - 251. Atlanta, Georgia: IEEE.

Kim, I. M., Shin, B. C., & Lee, D. J. (2000). SIR-based call admission control by intercell interference prediction for DS-CDMA systems. *IEEE Communications Letter*, *4* (1), 29-31.

Kim, K. I., & Kim, S. H. (2005). A Light Call Admission Control with Inter-Cell and Inter-Service Fairness in Heterogeneous Packet Radio Networks. *IEICE Transactions*, *88-B* (10), 4064-4073.

Kwon, T., Choi, Y., & Naghshineh, M. (1998). Optimal Distributed Call Admission Control for Multimedia Services in Mobile Cellular Network. *International Workshop on Mobile Multimedia Communication 1998 (MoMuC '98)* (pp. 477-482). Berlin: IEEE.

Kwon, T., Choi, Y., Bisdikian, C., & Naghshineh, M. (1999). Measurement-based Call Admission Control for adaptive multimedia in wireless/mobile networks. *IEEE Wireless Communications and Networking Conference 1999, (WCNC'99). 2*, pp. 540-544. New Orleans: IEEE.

Kwon, T., Kim, S., Choi, Y., & Naghshineh, M. (2000). Threshold-type call admission control in wireless/mobile multimedia networks using prioritised adaptive framework. *IEEE Electronics Letters*, *36* (9), 852-854.

Lai, F. S., Misic, J., & Chanson, S. T. (1998). Complete sharing versus partitioning: Quality of service managment for wireless multimedia networks. *7th International Conference on Computer Communications and Networks 1998* (pp. 584-593). Lafayette, Louisiana, USA: IEEE Computer Society.

Li, W., & Chao, X. (2007). Call Admission Control for an Adaptive Heterogeneous Multimedia Mobile Network. *IEEE Transactions on Wireless Communications*, *6* (2), 515-525.

Lindemann, C., Lohmann, M., & Thümmler, A. (2004). Adaptive call admission control for QoS/revenue optimization in CDMA cellular networks. *ACM Journal on Wireless Networks (WINET)*, *10* (4), 457-472 .

Liu, Z., & Zarki, M. E. (1994). SIR-based call admission control for DS-CDMA cellular systems. *IEEE Journal on Selected Areas in Communications*, *12* (4), 638-644.

Nasser, N., & Hassanein, H. (2004). Seamless QoS-Aware Fair Handoff in Multimedia Wireless Networks with Optimized Revenue. *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 1195-1198). Niagara Falls, Canada: IEEE.

O'Callaghan, M., Gawley, N., Barry, M., & McGrath, S. (2004). Admission Control for Heterogeneous Networks. *13th IST Mobile & Wireless Communications*. Lyon.

Racunica, I., Menouni Hayar, A., & Bonnet, C. (2004). An opportunistic scheduling with fairness for NRT traffic in presence of RT traffic for UMTS/TDD. *7th International Symposium on Wireless Personal Multimedia Communications 2004 (WPMC'04)*. Abano Terme, Italy: EURECOM+1410.

Rajaratmam, M., & Takawira, F. (1999). Performance analysis of highway cellular networks using generalised arrival and generalised service time distributions. *16th International Teletraffic Congress (ITC-16). 3*, pp. 11-22. Edinburgh: Elsevier.

Rajaratnam, M., & Takawira, F. (2000). Nonclassical Traffic Modeling and Performance Analysis of Cellular Mobile Networks with and Without Channel Reservation. *IEEE Transactions on Vehicular Technology*, *49* (3), 817-834.

Ramjee, R., Nagarajan, R., & Towsley, D. (1996). On optimal call admission control in cellular networks. *INFOCOM '96. 1*, pp. 43 - 50. San Francisco: IEEE.

Ramjee, R., Towsley, D., & Nagarajan, R. (1997). On optimal call admission control in cellular networks. *Wireless Networks*, *3* (1), 29-41.

Sen, S., Jawanda, J., Basu, K., & Das, S. (1998). Quality-of-Service degradation strategies in multimedia wireless networks. *48th IEEE Vehicular Technology Conference, 1998 (VTC 98)*, *3*, pp. 1884-1888. Ottawa.

Sherif, M. R., Habib, I. W., Nagshineh, M. N., & Kermani, P. K. (2000). Adaptive allocation of resources and call admission control for wireless ATM using generic algorithm. *IEEE Journal of Selected Areas in Communications*, *18* (2), 268-282.

Singh, S., Krishnamurthy, V., & Poor, H. V. (2002). Integrated voice/data call admission control for wireless DS-CDMA systems. *IEEE Transactions on Signal Processing*, *50* (6), 1483-1495.

Stasiak, M., Wisniewski, A., & Zwierzykowski, P. (2005). Uplink Blocking Probability for a Cell with WCDMA Radio Interface and Differently Loaded Neighbouring Cells. *Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources Conference/E-Learning on Telecommunications Workshop (AICT/SAPIR/ELETE'05)* (pp. 402-407). Lisbon: IEEE Computer Society.

Stratogiannis, D. G., Tsiropoulos, G. I., & Cottis, P. G. (2008). Call Admission Control in Wireless Networks: Probabilistic Approach and Efficiency Evaluation. *International Conference on Wireless Communications and Mobile Computing Conference 2008 (IWCMC '08)* (pp. 712-717). Crete Island: IEEE.

Tragos, E., Tsiropoulos, G., Karetsos, G., & Kyriazakos, S. (2008). Admission Control for QoS support in Heterogeneous 4G Wireless Networks. *IEEE Network Magazine*, *22* (3), 30-37.

Tsiropoulos, G. I., Stratogiannis, D. G., Kanellopoulos, J. D., & Cottis, P. G. (2008). Efficiency evaluation of class-based call admission control schemes for wireless communications. *IEEE International Symposium on Wireless Communication Systems (ISWCS '08)* (pp. 69-73). Reykjavik : IEEE.

Wang, X., Fan, F., & Pan, Y. (2008). A More Realistic Thinning Scheme for Call Admission Control in Multimedia Wireless Networks. *IEEE Transactions on Computers*, *57* (8), 1143-1147.

Warfield, R., Chan, S., Konheim, A., & Guillaume, A. (1994). Real-Time Traffic Estimation in B-ISDN Networks. *14th International Teletraffic Congress.* Antibes.

Wu, D. (2005). QoS provisioning in wireless networks. *Wireless Communications and Mobile Computing*, *5* (8), 957-969.

Yang, X., Feng, G., & Kheong, D. S. (2006). Call admission control for multiservice wireless networks with bandwidth asymmetry between uplink and downlink. *IEEE Transactions on Vehicular Technology*, *55* (1), 360-368.

Yavuz, E. A., & Leung, V. C. (2006). Computationally efficient method to evaluate the performance of guard-channel-based call admission control in cellular networks. *IEEE Transactions on Vehicular Technology*, *55* (4), 1412-1424.

Yilmaz, O., & Chen, I. (2006). Utilizing Call Admission Control to Derive Optimal Pricing of Multiple Service Classes in Wireless Cellular Networks. *12th International Conference on Parallel and Distributed Systems (ICPADS'06). 1*, pp. 605-612. Minneapolis: IEEE Computer Society.

Yu, O. T., & Leung, V. C. (1997). Adaptive resource allocation for prioritized call admission over anATM-based wireless PCN. *IEEE Journal on Selected Areas in Communication*, *15* (7), 1208 - 1225.

Communication Strategies for Strip-Like Topologies in Ad-Hoc Wireless Networks

Daniele De Caneva, Pier Luca Montessoro and Davide Pierattoni
University of Udine
Italy

1. Introduction

Many routing protocols have been designed for wireless sensor networks considering nodes that operate in a mesh topology. For specific application scenarios, however, a mesh topology may not be appropriate or simply not corresponding to the natural node deployment. Bridge (Kim et al., 2007) or pipeline (Jawhar et al., 2007) monitoring applications are examples where the position of sensor nodes is predetermined by the physical structure and application requirements. In this applications, where is clearly present a privileged dimension, it is quite natural to take advantage of it. Similar consideration can be made in more dynamic applications like the one of vehicular communication since the network can be approximated to have a linear topology without loss of accuracy.

This chapter will go through a description of the strategies developed so far to handle the problem of communication in strip-like topologies. For this specific problem several studies can be found in literature. Few research directions can be outlined: strip oriented routing, physical device design and specific MAC protocols. In the following four approaches are presented in order to describe how each direction can be investigated. The first two are related to the network layer of ISO/OSI protocol stack, the third one proposes use of devices with directional antennas while the fourth one designs a MAC protocol based on synchronous transmit-receive patterns. These approaches are somewhat complementary, each better suited for different scenarios.

2. Routing Layer Strategies

2.1 MERR

MERR (Minimum Energy Relay Routing) is a routing protocol which aims to address the problem of an economical use of power in wireless sensor networks. The goal is to minimize power consumption during communications in order to build networks for long-lasting operations. Its reference scenario is that of networks where sensors are deployed over a linear topology and have to send data to a single control center.

Assuming homogeneous sensor nodes deployed in an arbitrary linear sensor network, MERR permits every node to independently find a route to the base station that

approximates the optimal routing path. Finding a route means selecting appropriate relays between a sensor and the base station.

The problem of relaying data from nodes to the control center can be approached in two ways. The first is direct transmission, where every node transmits its packets directly to the base station. This approach suffers from important problems: first of all, in an environment with many obstacles or if the distance is too large, successful reception at the base station might not be feasible. Secondarily, with direct transmission, since the effort related with transmission increases as a power function of the distance, nodes far away from the base station will suffer greater power consumption and thus exhaust quickly their battery. From this considerations becomes clear that direct transmission is ideal only for scenarios where nodes are close to the base station or when the energy required for reception is large. In that case transmitting data directly to the control center, limits energy dissipation due to reception at the base station (which usually have unlimited power supply).

The second approach consists in taking advantage of the other nodes by using them as routers to forward data packets to the control center. MERR follows this method and in particular states the rules for router choice. MERR authors (Zimmerling et al., 2007) take distance from the MTE policy of routing where routers are chosen in order to minimize transmit energy. Minimizing transmit energy means choosing the nearest neighbor as router, with the evident drawback that a huge amount of energy is wasted if nodes are close to each other or the energy required for reception is high. MERR tries to respond to the question concerning which node must be chosen as router in order to obtain an energy efficient network. Zimmerling et al. based their work on that presented by Bhardwaj et al. (2001) where it is demonstrated that the optimal number of hops to reach a base station situated at a distance D is always:

$$K_{opt} = \left\lfloor \frac{D}{d_{char}} \right\rfloor \quad \text{or} \quad K_{opt} = \left\lceil \frac{D}{d_{char}} \right\rceil \tag{1}$$

where d_{char} is the characteristic distance, given by

$$d_{char} = \sqrt[n]{\frac{\alpha_1}{\alpha_2(n-1)}} \tag{2}$$

where α_1, α_2 and ε are parameters related to node's transceiver circuitry such that the power consumption involved in relaying r bit per second to a distance d meters onward (assuming a path loss of $1/d^n$) is

$$P_{relay}(d) = (\alpha_1 + \alpha_2 d^n)r \tag{3}$$

These results show that best performances are reached when packets perform $(K_{opt} - 1)$ relays by means of nodes equally spaced in intervals of D/K_{opt}.

Based on these assumptions, MERR states that every node should decide independently which will be its relay node. The choice is made seeking the down-stream node within the maximum transmission range whose distance is closest to the characteristic distance. After this decision is made by every node in the network, transmission power is independently reduced to the lowest possible level so that the radio signal can be received by the next-hop node without any errors. During normal functioning, a node will transmit data always to

the chosen relay node, regardless that this data comes from internal sensors or from another node.
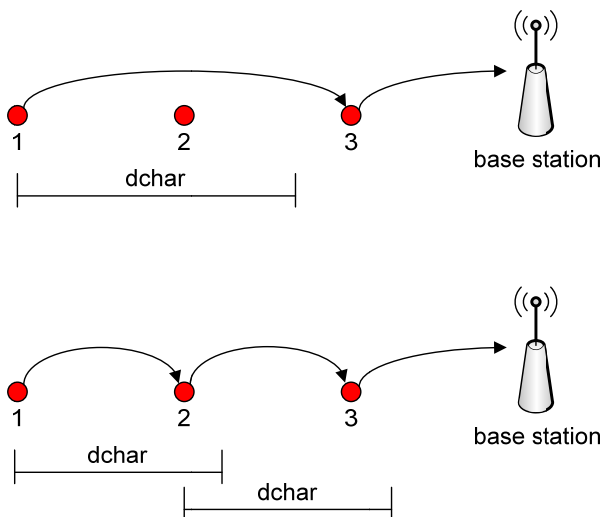

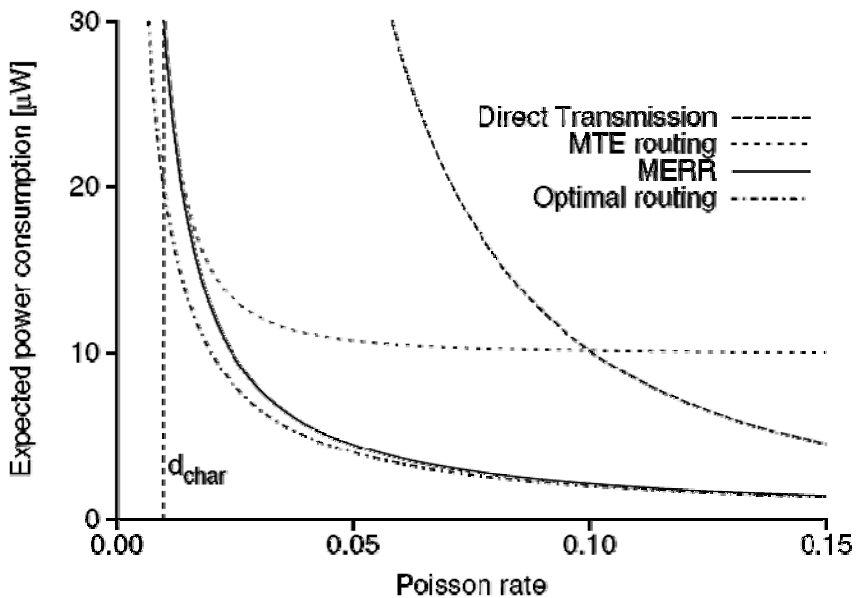Fig. 1. Characteristic distance influences packets routing path.

Fig. 2. Expected power consumption depending on Poisson rate λ for a constant number of sensors (n = 100) and path loss exponent 2.

In order to chose its own relay node, every sensor must know the characteristic distance (which is the same for all node if they are of the same kind) and the distance of all its neighbors (which can be manually measured during deployment or estimated using one of the methods present in literature such as Received Signal Strength or Time of Arrival).

Zimmerling et al. offer a comparison in terms of expected power consumption between MERR, optimal transmission, MTE and direct routing. For the sake of generality, the comparison is made using a one-dimensional homogeneous Poisson process with constant rate λ to model the distribution of nodes. The comparison, drown from a stochastic analysis made by the authors of MERR, clearly shows that energy consumption of MERR is always upper bounded by that of MTE. In particular MERR require less energy if the mean distance between nodes is lower than the characteristic distance.

2.2 Load Balanced Short Path Routing

Although not directly focused on strip-like topologies, the work presented by Gao et al. (2006) is worth mentioning because it covers the special case of a network where nodes are located in a narrow strip with width at most $\sqrt{3}/2 \approx 0.86$ times the communication range of each node.

Gao et al. tried to harness the main problem afflicting wireless networks, i.e. energy constraints. In particularly they focused on routing layer pointing out that, by minimizing path length, shortest path routing approaches minimize latency and overall energy consumption but may ignore fairness. In fact a protocol that searches the shortest path to route packets, will tend to abuse of some set of hops not exploiting all network resources. This behavior will quickly drain the batteries of involved nodes, causing the creation of holes within the network.

On the other hand load balanced routing strategies aim to use all available network resources in order to even the load, not regarding about communication performances.

Gao et al. in their work combined greedy strategies used to minimize path length and those used to evenly distribute load with the aim to achieve good performances in both metrics of latency and load balance. The problem of finding the most balanced routes is NP-hard even for a simple network and that is why Gao et al. firstly concentrated their efforts on a particular topology. The basic idea of their work is to maintain for each node a set of edges, called bridges, that are guaranteed to make substantial progress. In addiction their paper shows that, when a node has many neighbors, by distributing a collection of binary search trees on the nodes, memory needed on each node and routing/update cost can be reduced significantly.

The routing algorithm relies on two assumptions. The first is that each node knows its location, the second is that the rough location of the destination is known such that the source node knows whether it should send the packet toward its left or right.

For each node p, bc is a right (left) bridge if b and c are a couple of nodes visible to each other such that b is directly reachable by p, while c lies outside the communication range in a position that is right (left) to that of p (see Fig. 3). The load associated to the bridge is defined as maximum between the loads of b and c.

The routing is organized as follows: when p receives a packet, it first checks if the destination is a direct neighbor. In that case, it sends the packet to the destination. Otherwise, p chooses the lightest bridge, say bc, that forward the packet toward the

destination. Then p send the packet to b, where the process is repeated and so on till the destination is reached.

Gao et al provided a thorough demonstration that the algorithm works under the condition that strip width is equal or minor times the communication range of nodes. Additionally they presented simulation results over different network and traffic conditions.
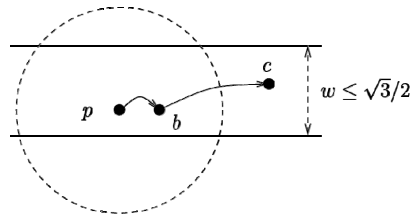

Fig. 3. Communication over a bridge.

3. MAC Layer Strategies

3.1 DiS-MAC

DiS-MAC (Directional Scheduled MAC) has been developed for wireless sensor networks that show a linear topology. It bases its functioning on a particular use of antennas. The premises that lead to this protocol is that directional or smart antennas have the potential to offer increased spatial reuse, longer communication ranges and the ability to point the radio beam toward a desired direction, properties that if properly exploited could potentially lessen the problem of interferences between nodes. Authors of DiS-MAC (Karveli et al., 2008) pointed out that current advances in antenna miniaturization techniques will open the doors of wireless sensor networks world to this kind of radiating systems.

The reference scenarios is that of highway and roadside monitoring sensors networks. Since roadsides and highways can be approximated to have a linear topology without loss of accuracy, Karveli et al. concentrated their effort on a sensor network deployed in such topology and consisting of N static nodes generating data packets of equal length with an arbitrary traffic rate. Every node is equipped with a directional antenna that can concentrate the main-beam to a particular direction and presents a some low gain side-lobes in other directions.
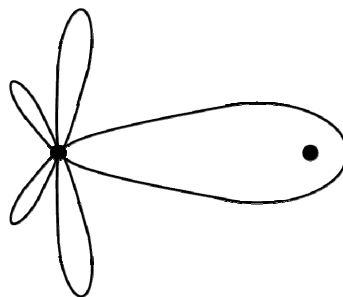

Fig. 4. Model of the antenna system radiation pattern.

Figure 4 shows the model for the radiation pattern used to develop the protocol. Other assumption for this protocol are that nodes are synchronized and that the traffic flows only in one direction.

Network synchronization permits to divide channel access in two phases of equal length. In the first phase every node occupying a odd position ($2n - 1$) directs its radiation beam in order to point to the subsequent node and then transmits its data. In this phase nodes occupying a even position ($2n$) switch their transceiver in reception mode. During the second phase roles are inverted: this time even nodes transmit data to their next node, while odd nodes perform reception. The alternation of phase I and phase II will continue indefinitely.

Phase I

2n-1   2n   2n+1   2n+2

Phase II

2n-1   2n   2n+1   2n+2
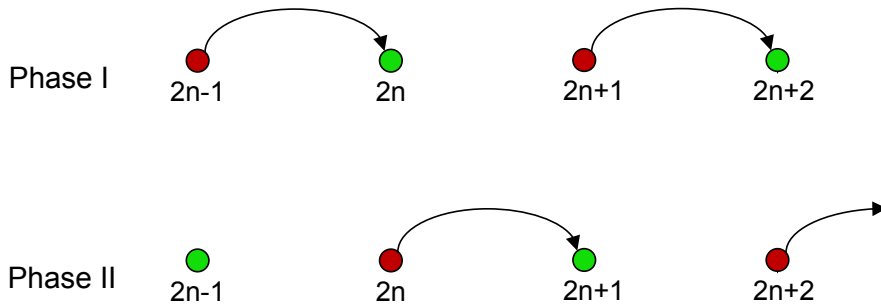
Fig. 5. Two phases scheduling.

This scheduled system provides a great efficiency, since it remove the possibility of collisions and the hidden terminal problem. In fact, since there is no contention, there is no need of control packets and thus it doesn't suffer from the overhead produced by them. This neatly configured system deterministically reaches a channel utilization equal to 1/2. This is quite impressive since in literature (Li et al, 2001) it is shown (both by simulations and experiments) that the capacity of a IEEE 802.11 network deployed in chain topology is limited to only 1/7. Additionally, thanks to the absence of channel contention, per hop latency, i.e. the time spent from packet generation at one node to its reception at the next node, is minimized and can be approximated by the duration of two phases.

Moreover the protocol is intrinsically robust because it limits interference between nodes, in fact when a node transmits, the first downstream node that can eventually suffer from this transmission is 3 hops ahead. Thus even considering the common assumption that the interference radius is twice the nominal transmission one, as shown in Fig. 6, DiS-MAC grants the avoidance of intra-network interference problems.

Authors of DiS-MAC outlined two extensions for their protocol. The first is a minor one, which states that if a node has no packet to transmit, it can enter into a sleeping mode. If another node have to transmit a packet to this sleeping node, it have to generate a short wake up radio signal in order to warn about the imminent transmission.

The second enhancement consist in the introduction of ACK packets to confirm that the transmitted packet has reached its destination without errors. Thanks to the contention-free nature of DiS-MAC, the repeated absence of ACK reception can be used as a marker of node failure. In this case, Karveli et al. have thought a strategy to react to the topology change. If node 2n fails, neither node $2n - 1$ nor node $2n + 1$ will receive its packets (the first one will

receive no ACK packet, while the second will receive no data packet). After a predefined counter expires, node   and node   will consider their neighbor failed and will start a recovery procedure. The first will extend its transmission range in order to reach the second one and then it will send a phase change request, which will propagated to all subsequent nodes. Phase change request is made through a special control packet and warns a node that its position into the chain is changed (e.g. node   is become node   ) and that it have to modify its behavior according to new topology. To avoid false failure detection, this protocol extension requires the transmission of periodical keep-alive packets.
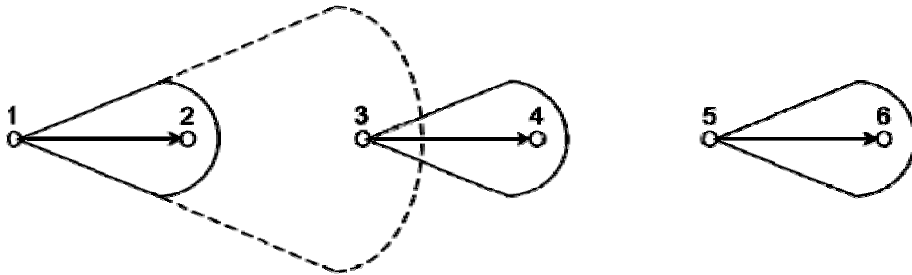
Fig. 6. Interference radius (dashed line) and transmit radius (solid line).

3.2 WiWi

The purpose of WiWi (De Caneva et al., 2008) is to emulate a wired link by means of an ad hoc network constituted by nodes distributed along a strip. The purpose of this wired link virtualization is that to handle scenarios where a single hop wireless link is not feasible and a wired link is not practical. An example could be given by a speleologist going deep down into the bowels of the earth, which can deploy the wireless network while it goes further with the exploration in order to maintain a communication channel with the outside world. Other examples can be found in all those situation where a multi-hop link is required. Moreover WiWi can be successfully used in monitoring applications.

Results presented in (Min & Chandrakasan, 2003) regarding power consumption over multi-hop networks demonstrated that the number of hops used to route a packet from source to destination should not be too high. In fact in such situation the portion of power consumption which is independent by the transmission distance became predominant, thus causing energy saving obtained by shorter transmission hops to be nullified. Moreover, the higher the hops number, the higher the latency. Nevertheless, the coverage range is limited by nodes' architectural characteristics and this define a lower bound for the hops number. This considerations led WiWi developers to choose a non-uniform node displacement and thus a cluster chain topology.
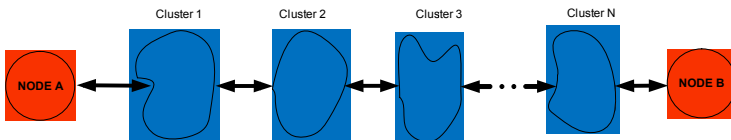

Fig. 7. WiWi topology.

De Caneva et al. made no assumptions over node deployment within the clusters, but full inter-cluster graph connection as well as complete radio coverage between nodes belonging to adjacent clusters.

WiWi protocol follows a synchronous full-duplex communication with fixed-side packets where clusters act as single nodes. In particular there exists two data stream which proceed along the chain in two different manners, depending on the direction. The first is a downward stream that relays packets from the head of the strip to the tail (gray packets in Fig. 8). This stream, which is responsible of maintaining network synchronization, follows a staggered pattern, i.e. a cluster sends a packet to the next cluster, which in turn immediately forwards it further down along the chain. This stream shows a latency equal to $L_{downlink} = hops\, T_s$, where T_s is the length of a time slot. The throughput associated with this stream can be expressed as the ratio between the number of bits forming a packet and the time interleaving two consecutive downstream transmissions (i.e. $\sigma\, T_s$, where σ is the number of slots by which spaces two consecutive transmissions).

The opposite stream follows the same principle of passing messages along the cluster chain, but between the reception of the packet and its forwarding, the cluster waits four time slots in order not to collide with the downward (Fig. 8 shows in different colors the steps taken by different upward packets). The latency affecting the upward stream is $\sigma - 1$ times the one of the downward, while the throughput is the same.

WiWi protocol is based on datagram transmission, in fact does not provide ACK packets to guarantee the correct packet exchange. Authors of WiWi point out that, if needed, the use of error correction codes could be introduced as well as acknowledgement mechanisms at higher level protocols.
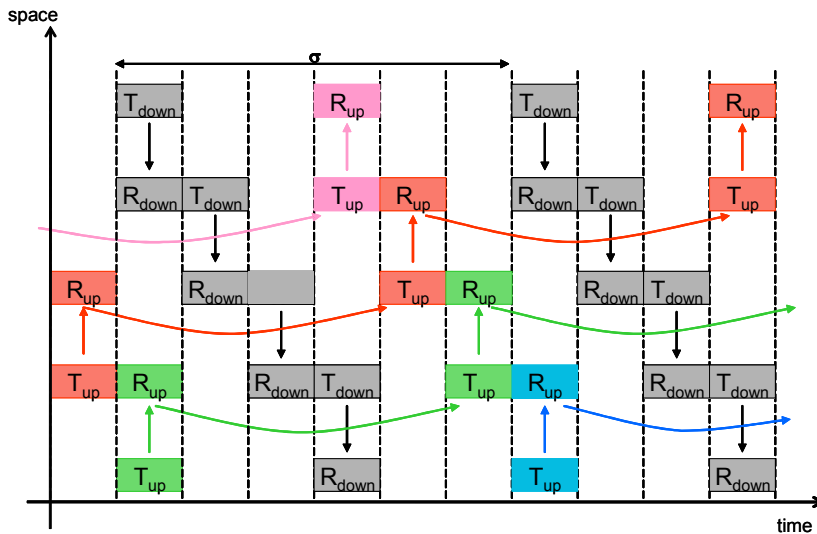

Fig. 8. Bidirectional, staggered transmission with symmetric throughput and asymmetric latency over a WiWi link.

As previously mentioned WiWi clusters act as a single node. This is done in order to reach redundancy as well as load balance. In fact WiWi requires that each cluster independently organize itself by ordering its nodes. By ordering a node belonging to the cluster can be elected as node on duty, i.e. the node that have to perform the packet relaying operations that compete to the cluster. The other nodes act as backup nodes. Operatively, during the reception slot every node of the cluster receives and stores the packet arriving from the previous cluster in the chain. In the subsequent transmission slot, the node on duty forwards the packet, while at the same time all backup nodes perform a sensing of the wireless channel. If the backup nodes perceive the loss of the duty node, they react autonomously by redefining their order within the cluster. This way the backup node, which would have its turn next to the current duty node, takes the role of forwarding the packet. The remaining backup nodes perform the sensing again in order to be sure that a backup node has reacted and the forwarding has occurred (Fig. 9). WiWi protocol grants an immediate redundancy equal to the number of backup nodes, which is the total number of active nodes in a cluster minus one. This means that the slot time upon which WiWi is based must have a duration capable to conserve this redundancy mechanism, which lead to a minor loss in throughput and latency performances. In the packet header could be inserted a notification flag to inform subsequent clusters of the failure event.

Clearly the node on duty is burdened with a higher power consumption, that is why nodes in turn cover this role following a round robin schedule. Additionally the scheduling of the duty evenly shares the load among cluster nodes extending the network lifetime and opening the door to the use of energy scavenging techniques.

The bandwidth unused by the redundancy mechanism, in normal conditions could be periodically exploited to reorganize each cluster on the run, in order to take care of the post-deployed nodes, if any.
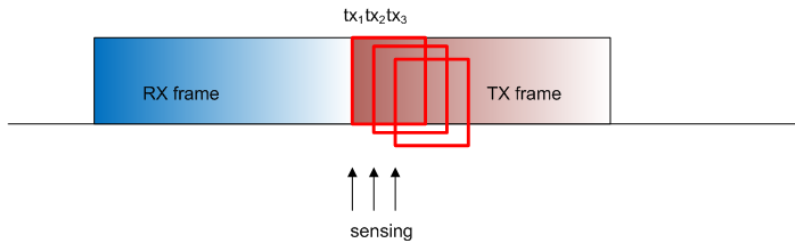

Fig. 9. Cluster redundancy management.

7. Conclusion

In this chapter were presented four algorithms whose aim is to manage packet relaying within an ad-hoc wireless network formed by nodes deployed over a strip. This algorithms are not exactly competing, instead they are focused on somewhat different scenarios which are related to different applications and hardware capabilities. In a field like the one of wireless sensor networks, where hardware constraints and application needs arise extremely challenging problems, taking every possible advantage is crucial. From this point of view it is clear that research have to develop new algorithms and protocols which exploit

network topology. Algorithms for linear and strip topologies represent the first steps toward this new trend of topology-oriented protocols.

8. References

Bhardwaj, M.; Garnett, T., Chandrakasan, A., "Upper bounds on the lifetime of sensor networks", *Proceedings of IEEE International Conference on Communications (ICC 2001)*, pp. 785–790, Jun. 2001.

De Caneva, D.; Montessoro, P.L.; Pierattoni, D., "WiWi: Deterministic and Fault Tolerant Wireless Communication Over a Strip of Pervasive Devices", *Proceedings of Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on*, pp.1-5, 12-14 Oct. 2008.

Gao, J.; Zhang. L., "Load-balanced short-path routing in wireless networks", *Parallel and Distributed Systems, IEEE Transactions on* , vol.17, no.4, pp. 377-388, April 2006

Karveli, T.; Voulgaris, K.; Ghavami, M.; Aghvami, A.H., "A Collision-Free Scheduling Scheme for Sensor Networks Arranged in Linear Topologies and Using Directional Antennas", *Proceedings of Sensor Technologies and Applications, 2008. SENSORCOMM '08. Second International Conference on*, pp.18 – 22, 25-31 August 2008.

Kim, S.; Pakzad, S.; Culler, D.; Demmel, J.; Fenves, G.; Glaser, S; Turon, M, Health "Monitoring of Civil Infrastructures Using Wireless Sensor Networks", *Proceedings of Information Processing in Sensor Networks, 2007. IPSN 2007. 6th International Symposium on*, pp. 254-263, 25-27 April 2007.

Jawhar, I.; Mohamed, N.; Shuaib, K., "A framework for pipeline infrastructure monitoring using wireless sensor networks", *Proceedings of Wireless Telecommunications Symposium, 2007. WTS 2007*, pp. 1-7, 26-28 April 2007.

Li, J.; Blake C.; De Couto D.S.J.; Imm Lee, H.; Morris, R., "Capacity of Ad Hoc Wireless Networks", *Proceedings of Mobile Computing and Networking, 7th ACM International Conference on*, pp. 61-69, July 2001.

Min, R. & Chandrakasan, A., "Top Five Myths about the Energy Consumption of Wireless Communication", *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol. 1, No. 2, 2003.

Zimmerling, M.; Dargie, W.; Reason, J.M., "Energy-Efficient Routing in Linear Wireless Sensor Networks", *Proceedings of Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE Internatonal Conference on*, pp. 1-3, 8-11 October 2007.

RSS Based Technologies in Wireless Sensor Networks

Samitha Ekanayake and Pubudu Pathirana
Deakin University
Australia

1. Introduction

Recent advances in electronics, computing and wireless communication technologies have made possible the use of low cost, low power sensor nodes with processing and wireless communication capabilities for variety of monitoring and control applications. A Wireless Sensor Network (WSN) is a collection of densely deployed such sensor nodes, having a collaborative objective. In typical WSN applications the positions of the sensor nodes are not engineered or predetermined. Instead the nodes are randomly deployed into the scenario. For example, in a large environmental monitoring sensor network (lake, forest, or seabed) involving thousands of sensor nodes, the nodes may be air dropped into the area of interest. In such WSN, the nodes are entirely dependent on the limited energy reserves such as batteries. Therefore nodal energy conservation is of utmost importance for prolonged network life. In this chapter we explore some RSS (Received Signal Strength) based techniques for power conservation in such randomly deployed WSN.

Although the WSN concept is being extensively explored in the recent past, there has not been an all-in-one communication scheme which satisfies the requirements of every networking scenario. We consider two networking scenarios which incorporate RSS based transmission power controlling to ensure Quality of Service (QoS) guaranteed communication links and to save limited nodal energy reserves. Both networking scenarios are having high application value in WSN arena, an all-to-all networking scenario and a mobile data collector based data collection network. In both networks we consider wireless nodes with multiple access communication capabilities (such as CDMA).

Among the multiple access schemes in wireless communications, CDMA has become the most promising technology that can satisfy most aspects in modern communication networks, such as higher speeds, larger client base and QoS guaranteed communication. Although CDMA started service in cellular communications in late 90's, the concept was originally introduced by Claude Shannon and Robert Pierce in 1949 (Ellersick 1984), and then extended by DeRosa-Rogoff, Price & Green and Magnuiki (Cooper and Nettleton 1978; Prasad and Ojanpera 1998). The early developments of this technology were primarily focused on the military and navigation applications (Batchelor, Ochieng et al., 1996). As the first civilian application, a narrow-band spread spectrum CDMA scheme for cellular communication was first proposed by Cooper and Nettleton in 1978 (Scholtz 1994) and then

developed to the IS-95 and CDMA2000 standards which are used in modern CDMA wireless communications (Knisely, Kumar et al. 1998).

Maintaining the Carrier-to-Interference Ratio (CIR), alternatively referred as the co-channel interference, at a desirable level is the main aspect of power control in CDMA networks. In CIR balancing, the transmission powers of every user device is controlled such that it ensures the co-channel interference of each link guarantees QoS reception. CIR balancing in a cellular system has two aspects: intra-cell CIR balancing and inter-cell CIR balancing. In intra-cell CIR balancing the user devices control the transmission power such that it provides a constant received power at the base station (Gilhousen, Jacobs et al. 1991) to avoid near-far problem. This method is currently in practice with CDMA standards such as IS-95 and CDMA 2000 (Schiller 2003). Inter-cell CIR balancing received widespread attention among the academic community after the problem reformulation by Zander et al. in (Zander 1992). This work was further investigated by Grandhi et al. (Grandhi, Vijayan et al. 1993; Grandhi, Vijayan et al. 1994; Grandhi, Yates et al. 1997) and the Distributed Power Control (DPC) scheme proposed by Foschini and Miljanic (Foschini and Miljanic 1993) has become a standard benchmark due to its academic and practical significance (see (Cai, Wang et al. 2004; Uykan and Koivo 2004; Uykan and Koivo 2006) for further improvements), which was later adopted into wireless communication standards.

Wireless sensor networks are inherently associated with restrictions in power consumption mainly due to the limited energy resources such as batteries. Therefore, unlike in cellular communications, the power control in wireless ad-hoc networks are basically focused on energy conservation. Many power conservation techniques introduced for such networks can be found in the past research literature (ElBatt and Ephremides 2004; Lim, Leong et al. 2005; Hou, Shi et al. 2006; Klein and Viswanathan 2006; Gomez and Campbell 2007). Among them routing optimization (ElBatt and Ephremides 2004; Hou, Shi et al. 2006; Klein and Viswanathan 2006) and transmission power control (Gomez and Campbell 2007) are the widely researched areas. However as opposed to the above, different effective methods such as sleep and wakeup procedures implemented in the hardware layer (Lim, Leong et al. 2005), were also proposed.

In the next section we discuss an all-to-all network for a wireless sensor network having multiple access communication capabilities. Such communication scheme is benificial for sharing of sensor data within the sensor network for real-time processing and decision making. The power control algorithm enable every node in the network to communicate with each other at the same time while consuming the minimum amount of energy for communication. In section 3, we introduce a transmission power control algorithm for a WSN having a mobile data collector based data gathering system. This scheme ensures maximum communication duration for nodes and the mobile data collector while using minimum possible energy for data communication.

2. All-to-all networking for instantaneous data sharing among the nodes

Recent past has witnessed a growing popularity in the multi-cast networking technologies, which have added advantages in the modern communication needs such as internet based multimedia services (news, distant learning etc), multimedia conferencing facilities for computers and mobile phones (Almeroth 2000; Chan, Modestino et al. 2007).

In multi-casting, the broadcasting of a single data packet to the network by the node dramatically improves the bandwidth usage in comparison to the unicast networks (one-to-one networks). In addition to the multimedia communication; distributed computing, parallel processing , swarm robotics , and wireless sensor networks where each node have some information to share with the other nodes have distinct advantages in employing all-to-all networks (multi-casting) (Chen, Chen et al. 1996).

All-to-all communications, proposed by Yang and Wang can be classified as: all-to-all broadcasting and all-to-all personalized exchange depending on the nature of the communication (Yang and Wang 2001). In the former case, the information (data packet) originating from a single node is propagated through the entire network and in the latter case every node has distinct information to share with every other node in the network. Routing algorithms for both network types have been extensively researched in the past (Akyildiz, Ekici et al. 2003; Guo and Yang 2006; Transier, Fubler et al. 2007). However, these routing algorithms were based on multi-hopping mesh and torus based network architectures and involve routing tree generation, forwarding link assignments, sub-network creation etc. They also have many practical difficulties in applying to all-to-all ad-hoc networks (Yang and Wang 1998; Yang 2006). In modern distributed / parallel processing applications, the network essentially consists of time varying nodes (location changes and addition / removal of nodes), which cause changes in the mesh / torus at each instance of architectural change. Moreover, those multi-hopping all-to-all networks comprises of hopping (routing) delays and increased network congestion with increasing network traffic, resulting in loss of vital information.

In this discussion, we consider a situation where an ad-hoc connected multiple-node wireless network requiring instantaneous all-to-all personalized communication, which is distributed within a close range such that the single-hop communication can be achieved between every node. The communication scheme introduced here enables all-to-all networking of the nodes without forwarding tree generation based on the spatial configuration of the nodes, i.e. node mobility, addition / removal of nodes etc. The proposed network uses CDMA based communication architecture which enables the entire network to communicate simultaneously. Moreover, we derive the capacity of the network in-terms of the number of nodes in the network and introduce a power control algorithm which ensures that all the nodes are transmitting at the minimum possible transmission power while maintaining the connectivity of the entire network ensuring interference free communication.

2.2 Problem Formulation

Now we formally introduce the power control problem together with the associated network architecture, control constraints and network capacity.

(a) Network Architecture:

Consider a single hop all-to-all wireless network (Ω) in which N nodes communicate with each other simultaneously (see Fig 1) using spread-spectrum multiple access protocol (such as CDMA). In this network, the nodes are broadcasting the data continuously, rather than maintaining node-to-node communication links. The broadcast data from a particular node, which is uniquely coded, can be accessed by every other node in the network.
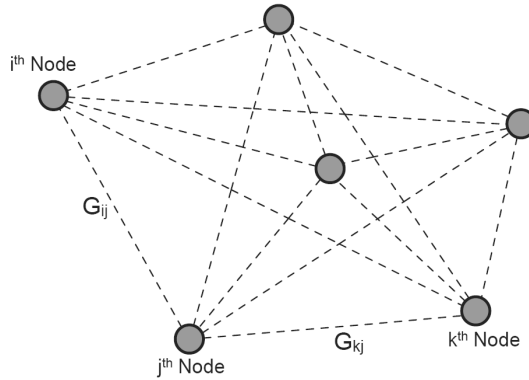
Fig. 1. - An all-to-all network consisting of six nodes.

The network model assumes followings;

- Nodes have instantaneous and error free Received Signal Strength (RSS) measurement capabilities.
- The measurements are immediately included in to the broadcast data, which will be used for the PC process.
- Link gain variations are negligible compared with the communication and the data processing time.
- All the nodes in the network are identical in performance (homogeneous).

In the controller analysis, the above assumptions are used in order to reduce the system complexity; however in later sections we relax these assumptions and present the controller behavior with erroneous measurements, non-homogeneous node properties, and link gain variations which resemble a real-world scenario.

(b) Control Constraints:

In order to achieve QoS guaranteed communication in every link, two conditions must be satisfied simultaneously; CIR constraint and the connectivity constraint.

CIR Constraint : Any node j in the network can receive the signal transmitted from any other node i, correctly, if the CIR measured at the jth node (γ_{ij}) is greater than the threshold CIR value γ_t. Then the CIR constraint can be defined as;

$$\gamma_{ij} = \frac{P_i G_{ij}}{\sum_{k \neq i, k \neq j}^{N} P_k G_{kj} + \eta} \geq \gamma_t, \forall i, k, j \in \Omega \qquad (1)$$

where P_i is the transmission power levels of the ith node. In the above expression, G_{ij} and G_{kj} are the link gains between i, j and k, j nodes respectively. Here the η represents the noise power (thermal noise) in the communication link and this is assumed to be constant for the geographical area (see (Foschini and Miljanic 1993; Uykan and Koivo 2004)).

Connectivity Constraint: To receive a signal from any node i, the received power level of the signal measured at the jth node (R_{ij}) must be greater than the receiver threshold R_{min},

which is the sensitivity of the receiver hardware. In this study the threshold received power is defined such that, the reception is not affected by the thermal noise of the band. Then the received power condition can be defined as (considering $R_{ij} = P_i G_{ij} + \eta$);

$$P_i G_{ij} + \eta \geq R_{min}, \forall i,, j \in \Omega. \tag{2}$$

(c) Capacity and spatial limitations:

In order to satisfy the above constraints, the all-to-all network has certain limitations in the spatial configuration and network capacity. This section derives the network capacity which ensure QoS guaranteed communication, and the relationships between the receiver sensitivity and the spatial configuration (link gains) to maintain reliable links.

From the connectivity constraint we get,

$$\min_{i,j \in \Omega} \left(P_i G_{ij} + \eta \right) \geq R_{min}, \tag{3}$$

which provides a condition that the network should satisfy at all the times for the power control algorithm to perform the desired action. Moreover, the network always satisfies the connectivity constraint (``connectivity guaranteed'' networks) if:

$$P_{min} G_{min} \geq R_{min} - \eta; \tag{4}$$

and the network is ``feasible'' if:

$$P_{max} G_{min} \geq R_{min} - \eta. \tag{5}$$

Here, P_{min} and P_{max} refers to the minimum and maximum transmission power levels of the nodes respectively, and G_{min} is the minimum link gain between any two nodes in the network. Above, the term ``feasible'' means that the network can achieve the connectivity constraint.

From the CIR constraint (equation (1));

$$\min_{i,j \in \Omega} \left(P_i G_{ij} \right) \left(\frac{1+\gamma_t}{\gamma_t} \right) \geq \max_{k,j \in \Omega} \left(\sum_{\substack{k \neq j}}^{N} P_k G_{kj} + \eta \right), \tag{6}$$

thus for ``connectivity guaranteed'' network:

$$P_{min} G_{min} \left(\frac{1+\gamma_t}{\gamma_t} \right) \geq (N-1) P_{max} G_{max} + \eta,$$

resulting,

$$N \leq 1 + \left(\frac{1+\gamma_t}{\gamma_t} \right) \left(\frac{P_{min} G_{min}}{P_{max} G_{max}} \right) - \left(\frac{\eta}{P_{max} G_{max}} \right). \tag{7}$$

For the ``feasible'' network:

$$P_{max} G_{min} \left(\frac{1+\gamma_t}{\gamma_t} \right) \geq (N-1) P_{max} G_{max} + \eta,$$

limiting the capacity as,

$$N \leq 1 + \left(\frac{1+\gamma_t}{\gamma_t} \right) \left(\frac{G_{min}}{G_{max}} \right) - \left(\frac{\eta}{P_{max} G_{max}} \right). \tag{8}$$

Definition 1. Limited Capacity Network: A multi-casting network satisfying the equation (8) on the number of nodes is defined as a Limited Capacity Network.

Remark: In above derivations, the network capacity is determined in terms of the number of nodes connected (N) at an instance and this number is dependent on the target CIR (γ_t). In spread spectrum networks, γ_t is selected to maintain the desired network quality, bandwidth and the data transfer speed (Gilhousen, Jacobs et al. 1991). Thus limiting the number of nodes to N ensures that the desired communication capacities/qualities are preserved in the network.

Remark: In limited capacity networks, the range of link gains in the desired geographical area ($G_{ij} \in [G_{min}, G_{max}]$) is a decisive factor on the number of nodes. However, this enables us to accurately select the number of nodes to be deployed in a particular region, knowing the range of link gain at that region.

Remark: In limited capacity networks, the maximum number of nodes (N_{max}) is defined such that the networks always satisfy the CIR constraints without directly depending on the spatial distribution of the nodes. However, this does not mean that a network with number of nodes $N > N_{max}$ in the same geographical area (not necessarily in the same configuration) does not satisfy the CIR constraints.

(d) Intended Controller Behavior for Energy Conservation:

In this power control problem, we consider an ad-hoc network satisfying ``*Limited Capacity*'' and "*feasible*" conditions. The problem considered here is to maintain all-to-all communication links in such networks, while minimizing the network power consumption via transmission power control. The proposed power control algorithm is focused on maintaining minimum requirements for satisfying the connectivity constraints, which automatically satisfies the CIR constraint in a Limited Capacity network.

2.3 Iterative Controller

In this section, we present a transmission power control scheme (see Fig 2) to maintain the received powers at the desired value that satisfy the connectivity of the network, and derive the tolerance limit for selecting the target received power.

The transmission power of the ith node (P_i) is determined by,

$$\dot{P_i} = a(e_i - R_t), \tag{9}$$

here $a < 0$ is a constant, e_i is the average received power at the other nodes, i.e

$$e_i = \frac{\sum_{j \neq i}^{N} \left(P_i G_{ij} + \eta \right)}{N - 1}$$

, and R_t is the target received power which satisfy the connectivity constraint for all the nodes. In this power control algorithm, we assume that the nodes are transmitting at the maximum transmission power at time zero (at the initialization of the algorithm).
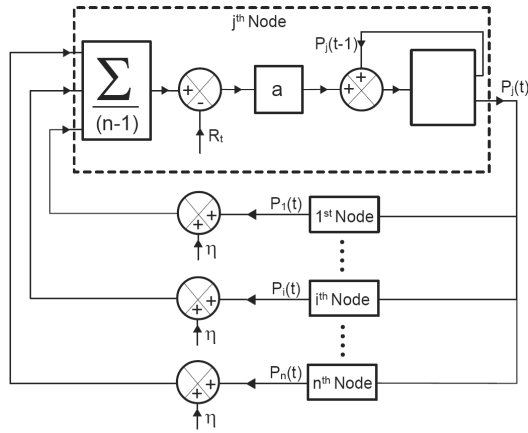
Fig. 2. - Block diagram representation of the controller of the jth node

(a) Convergence of the controller

From the definition we have, $e_i = P_i A_i + \eta$ and thus $\dot{e}_i = \dot{P}_i A_i$, where $A_i = \dfrac{\sum_{j \neq i}^{N} G_{ij}}{N-1}$ is the

``average link gain'' for the ith node.

With this, the controller function can be reformulated as:

$$\dot{e}_i = a A_i (e_i - R_t),$$

From the above expression it is evident that the control variable e_i converges to R_t, if $\| a A_i \| < 1$. Since $G_{i,j} < 0, \forall i, j$ and selecting $\| a \| < 1$ always satisfies the $\| a A_i \| < 1$ condition for the convergence.

(b) Satisfying Connectivity for Every Node

The convergence of the above controller describes the trajectory of the average received power, however, it does not say anything about the trajectories of the RSS in each link or their connectivity. In this section, we obtain a relationship between link gains, sensitivity of the receiver hardware and the target RSS value, which can be used to determine the tolerance limit when selecting R_t. This relationship is formally introduced in the following proposition.

Proposition 1: In an all-to-all network using the power controller described by (9) and deployed in a geographical area having link gains within a known range, i.e. $G_{ij} \in [G_{min}, G_{max}]$, the connectivity constraint is always satisfied if the threshold value for

the power controller, $R_t \geq \dfrac{R_m + \eta(X-1)}{X}$, where $X = \min_{i, j \in \Omega} \left(\dfrac{G_{ij}}{A_i} \right)$.

Proof. Let the error between the average RSS and the RSS of the node j,

$$e_{ij} = R_{ij} - e_i = P_i\left(G_{ij} - A_i\right)$$

and the time derivative;

$$\dot{e}_{ij} = \dot{P}_i\left(G_{ij} - A_i\right)$$

Then using the control function (9) we have,

$$\dot{e}_{ij} = aA_i\left[e_{ij} - (R_t - \eta)\left(\frac{G_{ij}}{A_i} - 1\right)\right]. \tag{10}$$

Above expression implies that e_{ij} converges toward $(R_t - \eta)\left(\frac{G_{ij}}{A_i} - 1\right)$, if the conditions for

the convergence of e_i are satisfied. Since the above statement is valid for any node i, j in

the network, we can determine the lower bound of e_{ij} as;

$$\min_{i,j\in\Omega}(e_{ij}) \le (R_t - \eta)\left(\min_{i,j\in\Omega}\left(\frac{G_{ij}}{A_i}\right) - 1\right). \tag{11}$$

For an all-to-all ad-hoc network deployed in the geographical area with $G_{ij} \in [G_{min}, G_{max}]$,

we have;

$$\min_{i,j\in\Omega}\left(\frac{G_{ij}}{A_i}\right) = \frac{(N-1)G_{min}}{G_{min} + (N-2)G_{max}}. \tag{12}$$

Then, the connectivity condition for any link i, j is satisfied if,

$$R_t + \min_{i,j\in\Omega}(e_{ij}) \ge R_{min},$$

i.e.

$$R_t \ge \frac{R_m + \eta(X-1)}{X} \tag{13}$$

where,

$$X = \min_{i,j\in\Omega}\left(\frac{G_{ij}}{A_i}\right)$$

which proves the assertion.

2.4 Simulation Results

(a) Power control algorithm: In this section a simulation case study is presented which illustrates the behavior of the system in an ideal situation described in the problem formulation section. The following parameters were selected for the simulation, $P_i \in [0.1, 3], G_{ij} \in [0.3, 0.6], R_{min} = 0.1, \gamma_{min} = 0.1$, $\eta = 0.05$, and $N = 4$.

With the selected parameters, the *feasible* condition ($R_{min} \le 0.3 \times 3 = 0.9$) and the ``Limited

Capacity'' network condition $\left(N \le 1 + \left(\frac{1.1}{0.1}\right)\left(\frac{0.3}{0.6}\right) - \frac{0.05}{0.6} = 6.417\right)$ are satisfied. The target

received power (R_t) is selected using equation (13) as: $R_t = 0.2 > \dfrac{0.1 + 0.05(0.6 - 1)}{0.6} = 0.1333$. The simulation results are shown in Fig 3. It is evident from the simulation figures that the controller converges to the minimum transmission power that satisfies all the constraints described in the previous section.
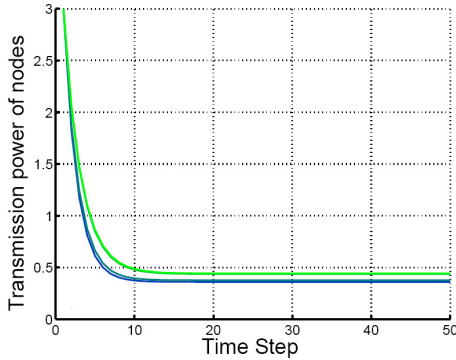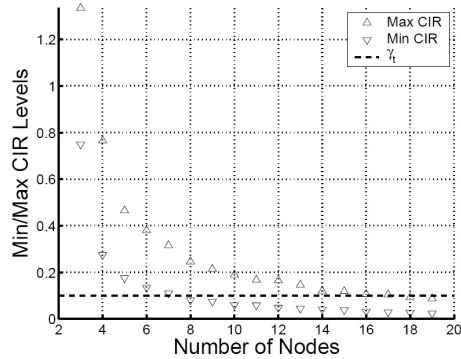

(a) Transmitted Powers
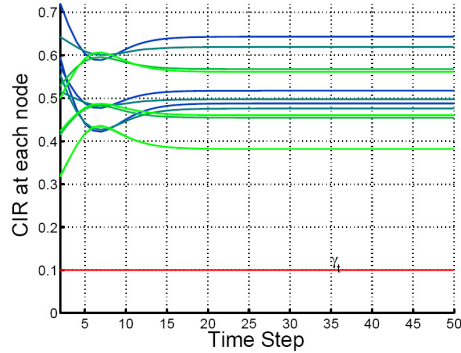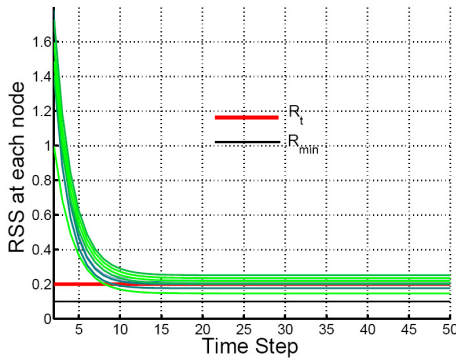
(b) CIR of each link

(c) RSSI of each link


(d) Behavior of CIR with number of nodes in the mesh

Fig. 3. - Effect of the transmission power control algorithm

(b) Network limitations: In this section we evaluate the theoretical assertions on the network limitations. In Fig 3.(d), the variation of CIR with increasing number of nodes is presented. In this figure, minimum and maximum CIR values for each node count are obtained by executing the simulation for 20 times with random selection of gain matrix , $G_{ij} \in [0.3, 0.6]$ and all other values are kept as in the previous case. It can be seen that the CIR range drops below the threshold value of 0.1 just after the node count exceeds 7 (the calculated maximum node count $N < 6.417$).

(c) Controller behavior in a real-world scenario: In this section, we illustrate the behavior of the proposed control scheme in the presence of real-world communication properties. Here the network is considered to have hertogeneous nodes, erroneous measurements and link gain variations (due to motion and other mobile obstacles).

The measurement error at j^{th} node is modeled as a normal distribution $v_j \in [0, \sigma_v^2]$, and the link gain is modelled as $\hat{G}_{ij} = G_{ij} 10^{(X/10)}$, where X is the dB attenuation due to shadowing effect and modelled as a Gaussian variable in the form of $X N(0, \sigma_X^2)$. Then the received power measurement can be modeled as,

$$R_{ij} = \left(P_i \hat{G}_{ij} + \eta \right) 10^{v_j/10}. \tag{14}$$
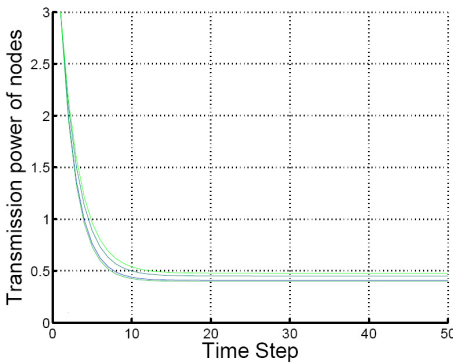
(a) Transmitted Powers
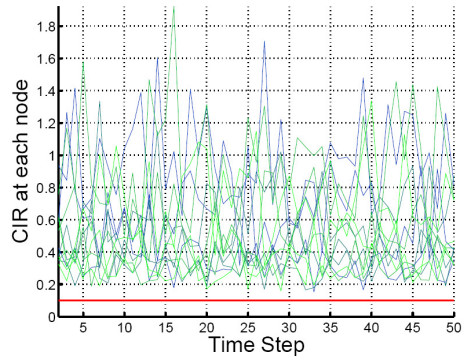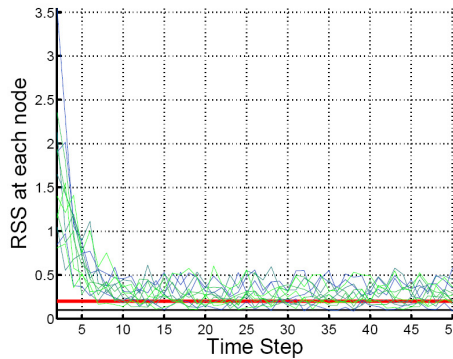
(b) CIR of each link

(c) RSSI of each link

Fig. 4. - Effect of the transmission power control algorithm - Real world scenario

The hetrageneous properties are modelled as differences in power transmission and power measurements, which are common in real-world communication equipments. The actual transmission power of the ith node is modelled as $\hat{P}_i = \alpha_i P_i$, where $\alpha_i \in (\alpha_{min},1)$ determines the power of the transmitter and is unique to each node. Similarly, the received power measurement performance factor $\beta_i \in (\beta_{min},1)$ determines the actual measured received power $\hat{R}_{ij} = \beta_i R_{ij}$.

In the simulation results shown in Fig 4, we consider $\sigma_X = 2$, $\sigma_v = 0.4$, $\alpha_{min} = 0.8$ and $\beta_{min} = 0.7$ (α_i and β_i values are randomly assigned for each node). According to the simulation results, the power control algorithm performs well in the presence of real-world limitations, maintaining the CIR of every link above the target (threshold CIR) as well as RSS of every link above the minimum RSS.

3. A simple power control algorithm for mobile data collector based remote data gathering scenario

In most low end networking devices CIR can not be directly measured, instead received power (in dBm) and Link Quality measurements can be obtained directly from the hardware. This raise the need of power control algorithms which do not entirely depending on CIR measurements, but depends on rather measurable parameters. In this study we derive the optimum value for co-channel interference measured at a base station, and introduce two power control algorithms to implement in user devices which can alter the transmission power to obtain the required CIR. In the proposed schemes we make use of Received Signal Strength (RSS) measurements to achieve the desired CIR at the base station.

3.1 Problem formulation
In this section we introduce the basic assumptions and models which will be used in the power control scheme. In this study we use the term ``Server'' to denote the base station node which communicate with all the ``Clients'' within the range. Here ``Client'' refers to the sensory node/ user device which is connected to the base station. The following notation is used throughout the paper.

P_i^T Transmission power of ith node. (dBm)

P_m^T Transmission power of server node. (dBm)

R_i^m Received power measured at client i, transmitted by server node. (dBm)

R_m^i Received power measured at server node, transmitted by node i. (dBm)

R_0^m Received power measured at reference distance (d_0), transmitted by server node. (dBm)

R_0^i Received power measured at reference distance (d_0), transmitted by client node i. (dBm)

Consider a wireless network with n clients connected to a single base station in a typical environment consisting of uncertainties in RF propagation due to shadowing, multi path propagation etc. The clients communicate with the server continuously using a common frequency band (as in CDMA). The server has a limit of n_{max} clients connected with it at an instant, and has a receive threshold of R_{min} (dBm) which is depending on the sensitivity of the receiver hardware. Also we assume that the communication network is not interfered by any other RF network in the domain of the base station (or ``cell'' in the cellular networking terminology). Throughout this paper, we assume that the server and the client maintain a continuous communication link, in which the server sends an acknowledgment signal back to the client for each data packet received (similar to (Uykan and Koivo 2004)). This signal contains the RSS of the received packet and the transmission power (if not transmitting at a fixed power) of the acknowledgment signal, which will be used in the PC algorithm. Also the communication hardware (server and client) have the capability to measure the RSS of each data packet.

3.2 Path loss model

We use the following path loss models for communication between the client node and the server node,

$$R_m^i = R_0^i - 10\eta \log_{10}\left(\frac{d}{d_0}\right) + S_{im} \qquad (15)$$

and

$$R_i^m = R_0^m - 10\eta \log_{10}\left(\frac{d}{d_0}\right) + S_{mi} \, , \qquad (16)$$

here term η refers to the path attenuation factor, which is a constant depending on the propagation media. In above expressions, S_{im} and S_{mi} refer to the combined effects due to shadow fading, multi-path propagation and any other fading effect occur from environmental factors such as presence of people, animals etc.

For line-of-sight communication in outdoor environments, specially long distance, the propagation of RF (Radio Frequency) waves can be approximated using free-space path loss model. This is possible as multi-path propagation and shadow fading effects do not become significant in such environments. Whereas, for wireless networks in indoor environments, the propagation is harder to predict due to presence of multi-path effects. Many researchers have studied the phenomenon of multi path propagation and proposed RSS models for indoor environments with the presence of obstacles (Tam and Tran 1995; Erceg, Greenstein et al. 1999; Santos, Alvarez et al. 2005; Sato, Sato et al. 2005; Puccinelli and Haenggi 2006). Applicability of those models for mobile nodes is debatable due to dynamic nature of environments and thus the model. Further, the experimental studies done by Lin et. al. in (Lin, Zhang et al. 2006) claims that the RSS value between two nodes in the line of sight have significant changes over the course of the day, thus location based mathematical models become inapplicable.
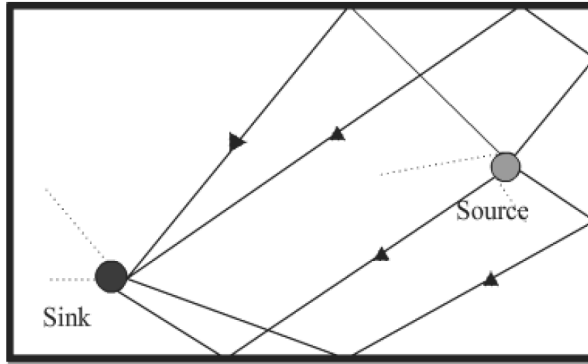
Fig. 5. - Modeling of 2D Multi-path propagation inside an enclosed environment, the figure presents few possible paths of multi-path propagation.

Ray-tracing concept for RF propagation, on the other hand, become a handy tool for predicting RSS variation in an indoor environment (Agelet, Fontan et al. 1997; Degli-Esposti, Lombardi et al. 1998; Remley, Anderson et al. 2000). Here, the radio waves are considered to follow the properties similar to visual light propagation in the presence of transparent obstacles. The effectiveness of Ray-Tracing method for RF waves increases with high frequencies. This is due to reduced scattering effects in shorter wave lengths. We use ray-tracing concept to make an assumption on the S_{xx} terms in equation (15) and (16), as follows;

$$S_{mi}(k) = S_{im}(k), \forall i = 1..n$$

where k represents a time step in discrete time.

As in Fig 5, there exists more than one path for receiving the signal from a RF source to a sink, and the overall S_{xx} term consists combination of all the multi-path propagation terms. With the assumption above, we claim that if the sink and source positions in the Fig 5 was interchanged then the only difference with the previous case is that the direction of propagation. That is all the multi-path links remain the same except the direction, thus results the same S_{xx} effect at the sink in the new configuration. The following experimental results justifies our assumption.

Basis for Ray-Tracing Assumption: In this experiment, four nodes (stationary nodes) were placed in an indoor environment and a mobile node communicating with them was randomly moved in the same environment. The received powers at each stationary node and the received power of the corresponding acknowledgment signal at mobile node were recorded. All the nodes are transmitted in a fixed transmission power and the corresponding R_0^m and R_0^i values were measured with $d_0 = 10cm$ (this was conducted in a large open space to minimize the effect of multi-path propagation). Then the expected value of the $S_{im} - S_{mi}$ can be written as follows.

$$E(S_{im} - S_{mi}) = E(R_m^i) - E(R_i^m) + E(R_0^m) - E(R_0^i)$$

The statistical data of the measurements and calculated $S_{im} - S_{mi}$ are presented in Table 1.

Parameter	Node 1	Node 2	Node 3	Node 4	Node 5
$E(R_m^i)/(dBm)$	-62.56	-65.96	-62.20	-62	-64.01
$E(R_i^m)/(dBm)$	-62.00	-64.00	-61.99	-61.94	-66.00
$E(R_0^i)/(dBm)$	-39.28	-40.65	-39.00	-41.00	-37.01
$E(R_0^m)/(dBm)$	-39.00	-39.00	-38.08	-41.00	-39.00
$E(S_{im} - S_{mi})/(dBm)$	0.29	0.31	-0.72	0.06	0.00

Table 1. - Expected values of measurements

From the experimental data it is evident that the $S_{im} - S_{mi}$ term is zero. In this experiment, even though all the transmitters are transmitting with the same power, we used the measured received powers at a reference distance rather than assuming $R_0^m = R_0^i$ in order to eliminate the effect of antenna gains.

In environments with such uncertainties (e.g. indoor, urban etc) ray-tracing concept can be used to predict the radio wave propagation (Degli-Esposti, Lombardi et al. 1998; Remley, Anderson et al. 2000). Here, the radio waves are considered to follow the properties similar to visual light propagation in the presence of transparent obstacles.

3.3 Power control analysis

(a) Optimum Carrier-to-Interference Ratio

CDMA base stations have a minimum CIR value (γ_{min}) which guarantee QoS reception. In CIR based power control algorithms such as (Foschini and Miljanic 1993; Uykan and Koivo 2004; Uykan and Koivo 2006) etc the controller is trying to maintain the CIR at a fixed value $\gamma_f \geq \gamma_{min}$. In this paper, we introduce a dynamic target CIR value ($\gamma^t \geq \gamma_{min}$) which is the optimal CIR for the number of clients connected with the server at that instance. The CIR, measured at the server, of the communication with the ith client (γ_i) can be defined as follows,

$$\gamma_i = \frac{R_i}{\sum\limits_{j=1, j \neq i}^{n} R_j} \qquad (17)$$

where R_i denotes the received power measured at the server, transmitted by the ith client in ``Watts''. Note that the R_i includes the random noise of the measurements as well. The server is said to have a good communication with the ith sensor, if the γ_i is greater than the threshold value γ^t. Then the above can be expressed in the following form (as in (Zander 1992)),

$$\frac{R_i}{\sum\limits_{j=1}^{n} R_j - R_i} \geq \gamma^t. \qquad (18)$$

The vector representation of the above is,

$$\left(\frac{1+\gamma^t}{\gamma^t}\right)\overline{R} \geq 1_n\,\overline{R}\,,\tag{19}$$

where 1_n is the unity matrix and $\left[\overline{R}\right]_i = R_i$. As proposed by Zander in (Zander 1992) we can derive the optimal γ^t value as follows (see Remark 1),

$$\gamma^t = \frac{1}{(n-1)}, \forall n < n_{max},\tag{20}$$

which results, $R_i = R^t, \forall i = 1\ldots n$, i.e. the received power values of the signals from every client, measured at the server should be equal. Here R^t is the target received power. This reduces the CIR balancing problem to a simple power control problem as presented in the next section.

Using the Perron-Froebenius theorem (see (Varga 1962)), the largest real eigenvalue of the matrix 1_n can be found as n. Selecting $R^t = R_{min}$ results in maintaining the CIR at the optimal value of $\dfrac{1}{(n-1)}$ while gaining the maximum energy saving in the network.

(b) Transmission Power Control

In this section, we propose a power control scheme to maintain the variable CIR presented above. Since we proved that maintaining a constant received power at the base station satisfies the optimal CIR condition, the ultimate target of the power control algorithm is to maintain R_m^i at R^t.

(c) Iterative Controller

The iterative power control algorithm is proposed as follows;

$$\dot{P}_i^T = f(R^i - R_m^i)\,.\tag{21}$$

Here the $f(\cdot)$ is defined as any function satisfying the Lipschitz condition,

$$f(|a-b|) \leq k_1\,|a-b|\tag{22}$$

where $k_1 \in [0,1]$ is the Lipschitz constant for the function $f(\cdot)$.

Proposition 1: The controller converges the R_m^i, starting from any arbitrary value, to R^t, if the transceiver gains remain constant.

Proof. From the path loss model between the client (15) and the server (16) nodes, we have

$$R_m^i = P_i^T - P_m^T + R_i^m)$$

and since P_m^T is a constant in our problem, the received power at the client node remains a constant. Then the controller becomes,

$$P_i^T = f(R^i - P_i^T + P_m^T - R_i^m)\tag{23}$$

resulting,

$$P_i^T = f(C - P_i^T + \upsilon_i)\,,$$

where $C = R^i + P_m^T - \hat{R}_i^m$ is a constant for the time interval. Here the υ_i is the random noise in the R_i^m, i.e. $\hat{R}_i^m = R_i^m + \upsilon_i$. Let $p = (C - P_i^T)$, then $\dot{p} = -\dot{P}_i^T$. The equation (23) can then be written in the vector form as,

$$\dot{\bar{p}} = -\phi(\bar{p} + \bar{\upsilon}) = \phi(-\bar{p} - \bar{\upsilon}) \, , \tag{24}$$

where $[\bar{p}]_i = P_i^T$, $[\bar{\upsilon}]_i = \upsilon_i$ and $\phi : \Re^n \rightarrow \Re^n$ i.e. $\phi(\bar{a}) = [f(a_1) \dots f(a_n)]^T$, $\bar{a} = [a_1 \dots a_n]^T \in \Re^n$ and $\phi(\bar{a}) = 0$ if $\bar{a} = 0$, thus the equilibrium point is the desired transmit power in (21) giving the optimal CIR in (20). Then as in (Uykan and Koivo 2004), selecting $\bar{a} = -\bar{p}_a - \bar{\upsilon}$ and $\bar{b} = -\bar{p}_b - \bar{\upsilon}$ yields,

$$\| \phi(\bar{p}_a) - \phi(\bar{p}_b) \| \quad \leq \quad k_1 \| (\bar{p}_a - \bar{p}_b) \| . \tag{25}$$

Since the above expression satisfies the Lipschitz conditions the system converges toward the desired power vector. (see (Uykan and Koivo 2004) and references there)

The numerical simulation results presented in Fig 6 shows the behavior of two controller functions; (1) A linear controller (f_L), and (2) A sigmoid based controller (f_S), defined as,

$$f_L(a) = 0.3 * a,$$

$$f_S(a) = 2\left(-0.5 + \frac{1}{1 + exp(-a)} \right)$$

Remark: Lipschitz constants of the $f_L(\cdot)$ is 0.3 and that of $f_S(\cdot)$ is 0.5 (see (Uykan and Koivo 2004)) thus the above control functions satisfy the condition in (22) and hence agree with the theoretical proof for convergence.
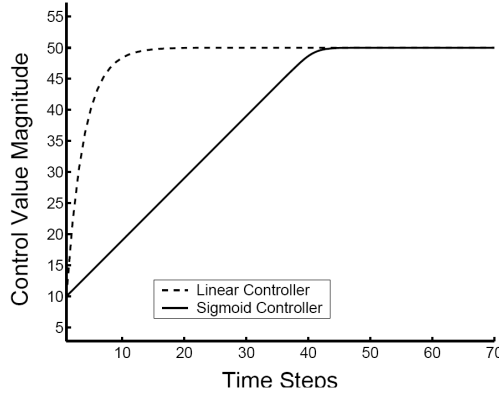
Fig. 6. - Numerical results showing the convergence of the controllers. Here $C = 50$ and $p(0) = 10$

3.4 Experimental Results

In the experimental evaluation we use two controller configurations, (i) Centralized implementation *(see Fig 7(a))* and (ii) Decentralized implementation *(see Fig 7(b))*. For the centralized implementation the server node transmits the signal strength of the received signal back to the client node, which will be used in the power control process. This uses the controller configuration expressed in the equation (21). In the distributed implementation, the client nodes make use of the local signal strength measurement for the power control process. For this approach the second configuration of the power control algorithm expressed by the equation (23) is used.

The experimental evaluation is conducted with the Micaz transceivers (Fig 8) developed by XBow technologies (Crossbow 2007). In Micaz hardware, the transmission power is controlled via an index (see (Chipcon 2004) on mapping of the index to dBm). The experiments were done for two basic cases, (i) static environment where the gains of the communication does not change significantly with in the time interval, and (ii) dynamic environment where the server node randomly moves within it's communication range. We use five cases for each environment to study the performance of the control algorithms. The controller implementation in each client node is shown in the Table 2.
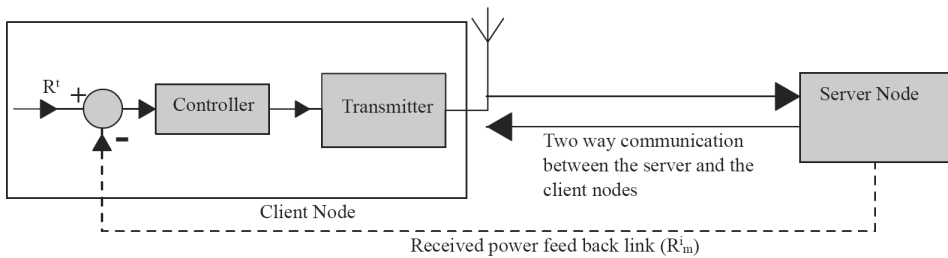
(a) Centralized Implementation of the controller
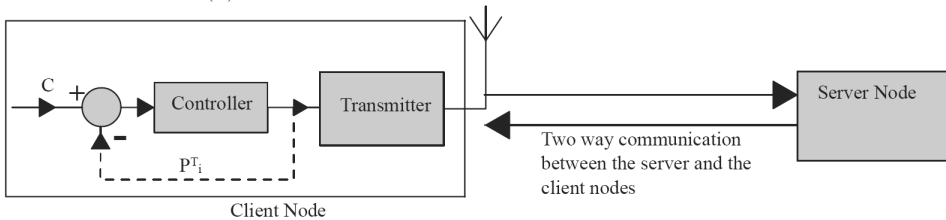
(b) Decentralized Implementation of the controller

Fig. 7. - Controller Configurations

Fig. 8. - Micaz node used for the experiment

(a) Static Environment

For this experiment we choose an environment with no or limited link gain variation (mostly due to the receiver noise). The Fig 9 shows the variation of received power measurements and the transmission power values of the client nodes. For this experiment, the target received power at the server node (R^t) is selected as $-70dBm$. According to the experiment results, the centralized controllers perform an accurate power control than the decentralized ones. Moreover, the centralized controllers demonstrate more robustness to measurement errors comparing with the decentralized one.

Client No.	Control Algorithm/ function
1	Centralized/ f_L
2	Centralized/ f_S
3	De-centralized/ f_L
4	De-centralized/ f_S

Table 2. - Client nodes and their controllers
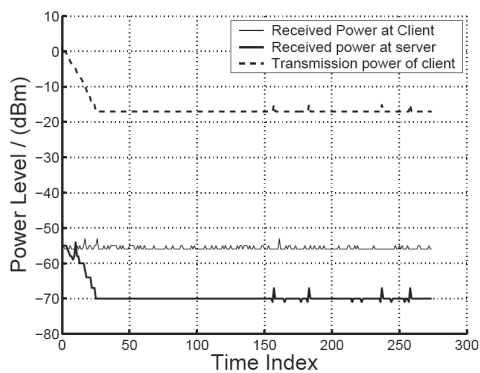
(a) Centralized implementation of the linear controller
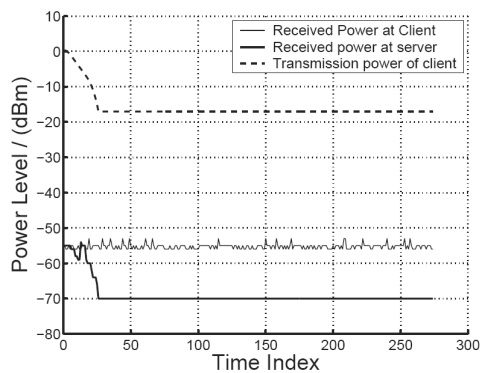
(b) Centralized implementation of the sigmoid controller
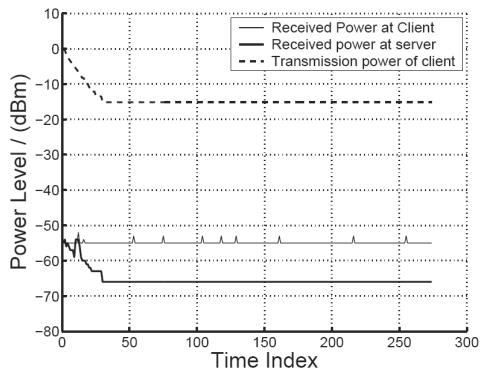
(c) De-centralized implementation of the linear controller
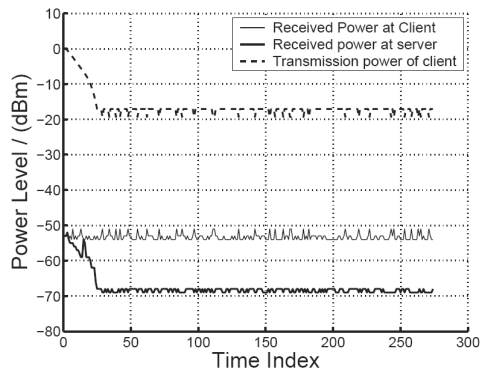
(d) De-centralized implementation of the sigmoid controller

Fig. 9. - Behavior of the iterative controller in a static environment
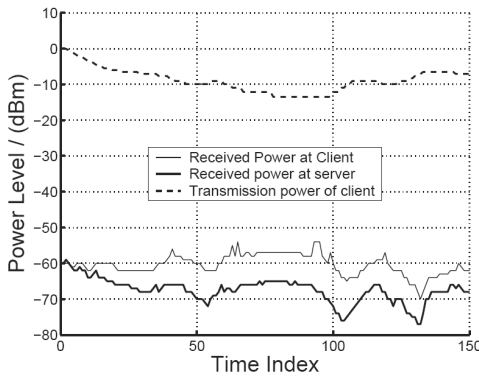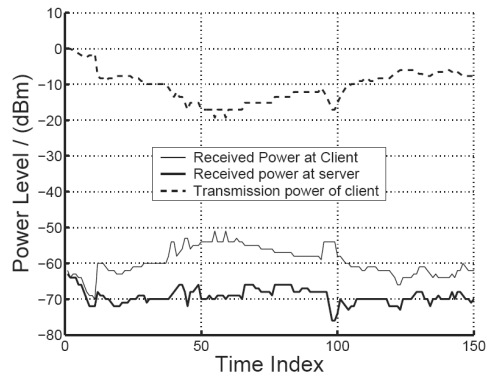
(b) Dynamic Environment

The Figure 10 shows the variation of received power measurements and the transmission power values of the client nodes. The target received power at the server node (R^t) is selected as $-70dBm$. In a dynamic environment, neither the centralized controllers nor the decentralized controllers perform well in maintaining a constant RSS at the server node. However, the centralized and decentralized implementation of the sigmoid function based controller performed well than the other controller configurations.

(a) Centralized implementation of the linear controller
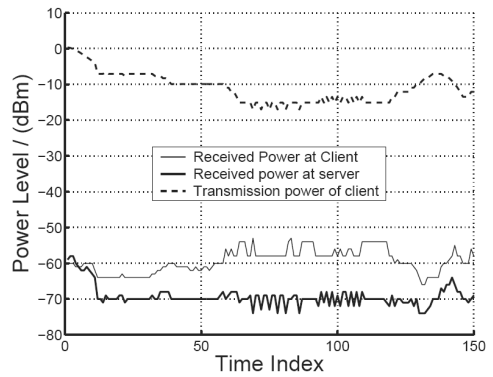
(b) Centralized implementation of the sigmoid controller

(c) De-centralized implementation of the linear controller

(d) De-centralized implementation of the sigmoid controller

Fig. 10. - Implementation of the iterative controller in a dynamic environment

4. Conclusion

The first section of this chapter introduces architecture for an all-to-all ad-hoc wireless network that satisfies the QoS requirements as well as power saving aspects. The CDMA based communication in the proposed network enables the operation in a very narrow band as well as maintaining a larger member base. This makes this network extremely suitable for military, swarm robotics and sensor network applications that require larger member base dispersed in relatively close proximity (i.e. within the single hop range of the transmitters) and simultaneous / delay-free communication within the network. The simulation case studies illustrate the behaviour of the controller in ideal conditions. Moreover, the theoretical assertions of network capacity and selection of target RSS value were illustrated.

Moreover, the controller behaviours in dynamic and real-world scenarios are tested using computer simulations.

In the second section of the chapter we introduced a power control algorithm which uses RSS measurements which is facilitated by most commercially available transceivers (in comparison with the CIR measurements presented in (Foschini and Miljanic 1993; Uykan and Koivo 2004) etc,). Since the control scheme focuses on maintaining the least power required for the base station / mobile data collector to capture the data packet, the clients transmit the signal in the minimum possible power which ensures the optimal CIR for every client. This effectively enhances the battery life of the power critical client nodes while maintaining a better quality of service. The experimental results verify the convergence of the power control scheme in a static environment as well as the practical applicability of the proposed controller.

5. References

Agelet, F. A., F. P. Fontan, et al. (1997). "Fast ray tracing for microcellular and indoor environments." Magnetics, IEEE Transactions on 33(2): 1484-1487.

Akyildiz, I. F., E. Ekici, et al. (2003). "A Distributed Multicast Routing Scheme for Multi-Layered Satellite IP Networks." Wireless Networks 9(5): 535-544.

Almeroth, K. C. (2000). "The evolution of multicast: from the MBone to interdomain multicast to Internet2 deployment." Network, IEEE 14(1): 10-20.

Batchelor, A., W. Y. Ochieng, et al. (1996). Design features of D-GNSS reference stations. Satellite Systems for Mobile Communications and Navigation, 1996., Fifth International Conference on.

Cai, M., W. Wang, et al. (2004). Power control algorithm for time-varying CDMA cellular systems. Intelligent Mechatronics and Automation, 2004. Proceedings. 2004 International Conference on.

Chan, Y. S., J. W. Modestino, et al. (2007). "An End-to-End Embedded Approach for Multicast/Broadcast of Scalable Video over Multiuser CDMA Wireless Networks." Multimedia, IEEE Transactions on 9(3): 655-667.

Chen, M.-S., J.-C. Chen, et al. (1996). "On general results for all-to-all broadcast." Parallel and Distributed Systems, IEEE Transactions on 7(4): 363-370.

Chipcon (2004). "C2240 Transceiver Datasheet." Chipcon Systems.

Cooper, G. R. and R. W. Nettleton (1978). "A spread-spectrum technique for high-capacity mobile communications." Vehicular Technology, IEEE Transactions on 27(4): 264-275.

Crossbow (2007). "Crossbow Technologies Incorporated." http://www.xbow.com.

Degli-Esposti, V., G. Lombardi, et al. (1998). "Measurement and ray-tracing prediction of indoor channel parameters." Electronics Letters 34(22): 2167-2168.

ElBatt, T. and A. Ephremides (2004). "Joint scheduling and power control for wireless ad hoc networks." Wireless Communications, IEEE Transactions on 3(1): 74-85.

Ellersick, F. (1984). "A conversation with Claude Shannon." Communications Magazine, IEEE 22(5): 123-126.

Erceg, V., L. J. Greenstein, et al. (1999). "An empirically based path loss model for wireless channels in suburban environments." Selected Areas in Communications, IEEE Journal on 17(7): 1205-1211.

Foschini, G. J. and Z. Miljanic (1993). "A simple distributed autonomous power control algorithm and its convergence." Vehicular Technology, IEEE Transactions on 42(4): 641-646.

Gilhousen, K. S., I. M. Jacobs, et al. (1991). "On the capacity of a cellular CDMA system." Vehicular Technology, IEEE Transactions on 40(2): 303-312.

Gomez, J. and A. T. Campbell (2007). "Variable-Range Transmission Power Control in Wireless Ad Hoc Networks." Mobile Computing, IEEE Transactions on 6(1): 87-99.

Grandhi, S. A., R. Vijayan, et al. (1994). "Distributed power control in cellular radio systems." Communications, IEEE Transactions on 42(234): 226-228.

Grandhi, S. A., R. Vijayan, et al. (1993). "Centralized power control in cellular radio systems." Vehicular Technology, IEEE Transactions on 42(4): 466-468.

Grandhi, S. A., R. D. Yates, et al. (1997). "Resource allocation for cellular radio systems." Vehicular Technology, IEEE Transactions on 46(3): 581-587.

Guo, S. and O. Yang (2006). "A Constraint Formulation for Minimum-Energy Multicast Routing in Wireless Multihop Ad-hoc Networks." Wireless Networks 12(1): 23-32.

Hou, Y. T., Y. Shi, et al. (2006). "Maximizing the Lifetime of Wireless Sensor Networks through Optimal Single-Session Flow Routing." Mobile Computing, IEEE Transactions on 5(9): 1255-1266.

Klein, T. E. and H. Viswanathan (2006). "Centralized power control and routing policies for multihop wireless networks." Information Theory, IEEE Transactions on 52(3): 849-866.

Knisely, D. N., S. Kumar, et al. (1998). "Evolution of wireless data services: IS-95 to cdma2000." Communications Magazine, IEEE 36(10): 140-149.

Lim, S., K. M. K. H. Leong, et al. (2005). "Adaptive power controllable retrodirective array system for wireless sensor server applications." Microwave Theory and Techniques, IEEE Transactions on 53(12): 3735-3743.

Lin, S., J. Zhang, et al. (2006). ATPC: adaptive transmission power control for wireless sensor networks. Proceedings of the 4th international conference on Embedded networked sensor systems. Boulder, Colorado, USA, ACM Press: 223-236.

Prasad, R. and T. Ojanpera (1998). A survey on CDMA: evolution towards wideband CDMA. Spread Spectrum Techniques and Applications, 1998. Proceedings., 1998 IEEE 5th International Symposium on.

Puccinelli, D. and M. Haenggi (2006). Multipath fading in wireless sensor networks: measurements and interpretation. Proceeding of the 2006 international conference on Communications and mobile computing. Vancouver, British Columbia, Canada, ACM Press: 1039-1044.

Remley, K. A., H. R. Anderson, et al. (2000). "Improving the accuracy of ray-tracing techniques for indoor propagation modeling." Vehicular Technology, IEEE Transactions on 49(6): 2350-2358.

Santos, R. A., O. Alvarez, et al. (2005). Experimental analysis of wireless propagation models with mobile computing applications. Electrical and Electronics Engineering, 2005 2nd International Conference on.

Sato, R., H. Sato, et al. (2005). A SBR estimation for indoor wave propagation through dielectric walls. Antennas and Propagation Society International Symposium, 2005 IEEE.

Schiller, J. (2003). Mobile Communications, Addison-Wesley.

Scholtz, R. A. (1994). The evolution of spread-spectrum multiple-access communications. Spread Spectrum Techniques and Applications, 1994. IEEE ISSSTA '94., IEEE Third International Symposium on.

Tam, W. K. and V. N. Tran (1995). "Propagation modelling for indoor wireless communication." Electronics & Communication Engineering Journal 7(5): 221-228.

Transier, M., H. Fubler, et al. (2007). "A hierarchical approach to position-based multicast for mobile ad-hoc networks." Wireless Networks 13(4): 447-460.

Uykan, Z. and H. N. Koivo (2004). "Sigmoid-basis nonlinear power-control algorithm for mobile radio systems." Vehicular Technology, IEEE Transactions on 53(1): 265-270.

Uykan, Z. and H. N. Koivo (2006). "Proportional power control algorithm for time varying link gains in cellular radio systems." Vehicular Technology, IEEE Transactions on 55(1): 341-349.

Varga, R. S. (1962). Matrix Iterative Analysis. N.J., Prentis Hall, Inc.

Yang, L.-L. (2006). "MIMO-assisted space-code-division multiple-access: linear detectors and performance over multipath fading channels." Selected Areas in Communications, IEEE Journal on 24(1): 121-131.

Yang, Y. and J. Wang (1998). "On blocking probability of multicast networks." Communications, IEEE Transactions on 46(7): 957-968.

Yang, Y. and J. Wang (2001). "Pipelined all-to-all broadcast in all-port meshes and tori." Transactions on Computers 50(10): 1020-1032.

Zander, J. (1992). "Performance of optimum transmitter power control in cellular radio systems." Vehicular Technology, IEEE Transactions on 41(1): 57-62.

Smart wireless communication platform IQRF

Radek Kuchta, Radimir Vrba and Vladislav Sulc
Brno University of Technology, Microrisc s. r. o.
Czech Republic

1. Introduction

Wireless communications systems are used in different areas of human activity. Wireless communications can be distinguished between licensed and non-licensed, according to the applied frequency band. Non-licensed bands are different in a lot of countries. In European Union, there are 433 MHz, 868 MHz, 2.5 GHz and other bands. In the United States of America, there are 916 MHz and others. These frequencies are very often used for interconnection of sensors, actuators, equipments, controllers, computers, remote controllers etc. They are used at least in two basic lines of work. The first one is for home automation and the second one is for industrial automation. Communications standards and communication protocols exist in both of these lines.

One such standardized protocol is, for example, Zigbee. It involves a solution based on the IEEE 802.15.4 standard (De Nardis and Di Benedetto 2007) prepared by Zigbee Alliance (ZigBee 2009). Among the proprietary solutions, reference can be made to the technology of MiWi launched by Microchip Technology Inc. (Flowers and Yang 2008), based on the aforementioned standard but simpler than Zigbee from the point of view of implementation and not allowing direct cooperation with Zigbee devices (Huang et al. 2008; Ji et al. 2008; Song and Yang 2008). Among the other solutions available on the market, mention would be made, for example, of the solution promoted by Z-wave alliance.

These solutions have disadvantage in attempt on being a universal solution targeting every kind of applications. It brings heavier protocols, more difficult and more expensive implementations.

Implementation of solutions such as Zigbee or MiWi consists of software solution stack and hardware solution used for communication. Software solution stack is developed by a microcontroller manufacture for defined microcontroller or by a producer that wants to supply his products for communication modules designed for the area of domestic automation. The software stack is a package of program routines, functional components and program subsystems (hereinafter Stack) permitting the basic operation of the communication module according to the chosen solution for wireless communication. The manufacturer of the end device uses the modules for the selected communication solution, and then, it creates a further application extension to implement the actual application functionality of the end device (Ferrari et al. 2007; Ghazvini et al. 2008; Chan 2008; Liang et al. 2008).

The need for this step is obvious from the point of view of the manufacturer of the processor products – the manufacturer wants to supply his products within the framework of the whole communication solution, and because he has the best knowledge of his own products, he creates the Stack referred to above, generally in much less time than each individual manufacturer of end devices would have taken to create his own product. By creating a Stack, he thus makes it possible to participate on communication solutions based on the processors supplied by itself, and it offers them not just to one but to many potential customers.
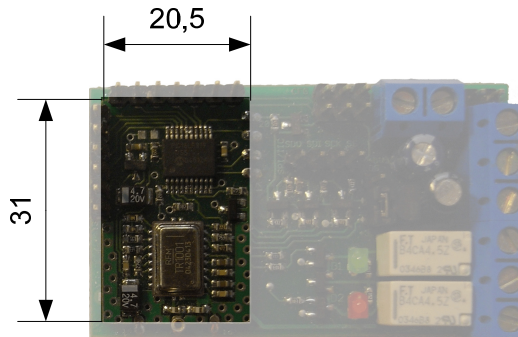
Fig. 1. The first generation of IQRF communication modules

Wireless communications platforms based on the standard communication networks, such as Wi-Fi or Bluetooth, are available on the market place. It is possible to buy a communication module with a simple communication interface that implements all needed functions and protocols. It is not so difficult to implement these modules and networks to the new devices. These solutions are useful for fast communications with greater volume of data. These modules have usually higher energy consumption, so they are not targeting low power applications.
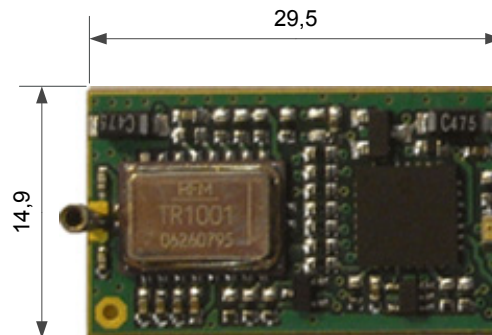
Fig. 2. The second generation of IQRF communication modules

There are also available proprietary solutions like Z-wave (Z-Wave 2009), radio modems XTR-434L (Aurel 2008), nRF24xx (Leonard 2007), RC1280HP (Vojacek 2007), etc. They

usually use master/slave communication model. Sometimes, they offer other integrated peripherals as AD converters, LEDs, or digital inputs/outputs.

Z-wave, for example, uses a mesh network topology with no master node. Any device can originate the message. If the preferred route is unavailable, the message originator will attempt other routes until a path is found to the recipient node. Z-wave rating units cannot be in sleep mode.

The chapter is focused on proprietary wireless communication platform IQRF. The platform supports different network topologies, allows fast and easy implementation to the new applications without deeper knowledge of the issue of wireless communications.

At the beginning of the chapter main features and hardware parameters of the IQRF platform are described. The next section contains description of the IQRF operating system with basic functionality description. Then IQRF gateways and available development tools are discussed. The next section contains description of IQMESH communication protocol used by IQRF platform. At the end of the chapter is a future work description and short summary.

2. Wireless communication platform IQRF

The IQRF platform was designed to address smaller segments of wireless market - buildings automation and telemetry. The platform was developed by Microrisc company (Microrisc 2009b). Main parts of the platform are covered by Czech and US patents (Sulc 2007a; b; c; 2008). These patents cover a method of creating a generic network communication platform, special signal coding scheme, and direct peripheral addressing in wireless network.
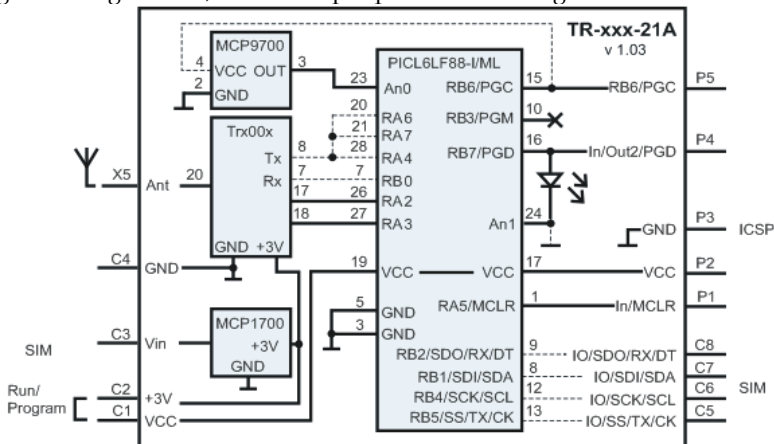

Fig. 3. The block structure of the IQRF module (Microrisc 2008a)

The IQRF platform is based on second generation of short-range radio components produced by RFM Company (RFM 2009). It works in non-licensed communication bands. IQRF communication modules (Microrisc 2008b) are available for 868 MHz and 916 MHz frequencies. Basic features of this wireless communications platform are especially very low power consumption, network possibility, programmable RF power up to 1.3 mW, optionally up to 10 mW, 170 m range, and 15 kb/s RF bit rate, optionally 100 kb/s. The first generation

of IQRF module is in Fig. 1, the second generation is in Fig. 2. Sizes in figures are in mm. The pictures aren't in scale. The second generation module is the same size like SIM card and used the same connector for interconnection with other parts of system. The block structure of the IQRF module is shown in Fig. 3.

Basically, the IQRF communication module has three basic input/output interfaces, one analog input, an SPI interface, and digital ports. Each module contains integrated analog temperature sensor, LED and 3 V linear regulators, which can be used for user application.

2.1 IQRF operating system

IQRF communications modules have own operating system. SW developers don't need to implement any part of wireless communication protocol. They only use prepared functions of operating system for their application. Whole system offers about 40 functions. A function block diagram is shown in Fig. 4. The main functions of OS are:

- RF functions for transmitting, receiving, bonding and setting up,
- IIC and SPI communication functions,
- EEPROM access functions,
- three buffers for RF, COM and INFO are available,
- some other auxiliary functions for LED, OS information, delays and sleep mode functions are available too.

Up to 32 bytes is possible to send in one packet.
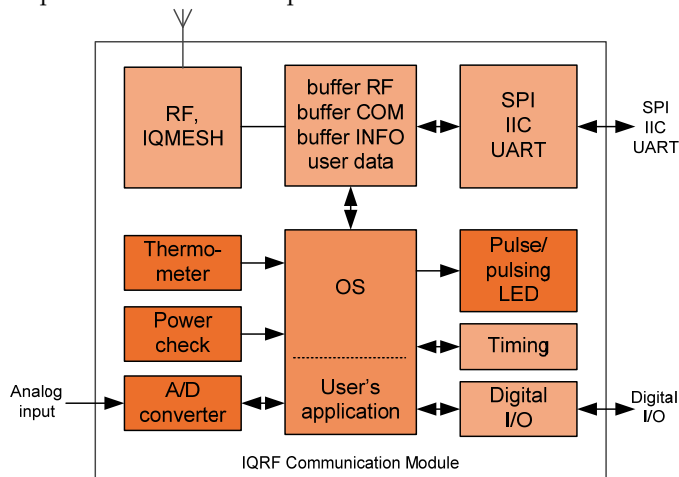
Fig. 4. Basic functionality block diagram of IQRF Operating system

IQRF operating system is implemented to the program memory of the microcontroller. Program memory is divided to two main parts. The first part is used by IQRF operating system and the second is available for user's application. When user's application needs to call some OS function, it calls function address defined in the definition file of the selected OS version. Programmers of the application can use whole set of the microcontroller instruction. Some restrictions for direct program memory access are applied. Because direct program memory access instructions are not allowed in the user's code, IQRF has

implemented functions to store and read data from the on chip integrated EEPROM memory.

IQRF is wireless communication platform, so IQRF OS support functions to create network, with different topology. When IQRF networking functionality is used, it network exist coordinator and unit. They have very similar OS, differences are in the function to control network that are implemented only in the coordinators modules.

To support wireless and network functionality tree data buffers are available. The OS also offers functions to copy data between buffers. Buffer called RF contains wirelessly received data or data to be transmitted. COM buffer is used to send and receive data via SPI, IIC and UART interface. INFO buffer is used by system for block operations.

OS also offers functions for timing, power control, reset and integrated LED control. Detailed description of all IQRF OS function is in (Microrisc 2008b).

2.2 IQRF gateways and development tools

Various gateways to common standards, such as Bluetooth, ZigBee and GSM are available. Simple applications can use RS-232 gateway or more useful USB gateway. These simple gateways were developed to allow connection between IQRF and other proprietary solutions. They also allow connecting IQRF and standard PC with user's application.

For more sophisticated applications, GSM or Ethernet gateways are available. To allow interconnection between IQRF and standard wireless solution a Bluetooth and ZigBee gateways are available.

Development tools allow debugging and testing of user applications using supporting software. To provide comfortable environment for a transceiver development kits typically contain interface connectors, battery, interface to user pins and so on.

3. IQMESH

IQMESH (Intelligent Mesh) protocol was defined in 2005 as a basic communication protocol for IQRF device with target to address mainly low power, low data rate, small wireless applications, like a home automation, office automation and telemetry (Microrisc 2008b).

IQRF utilize several unique and patented features, IQMESH protocol was defined to support them.

For instance, the patented method of creating a generic network communication platform with transceivers defines the simultaneous work of devices in two or more wireless networks allowing network chaining (Sulc 2007b). Example of IQMESH network chaining is shown in Fig. 5.

Two networks in Fig. 5, Network 1 and Network 2, are independent IQRF wireless networks. Every such network has one Coordinator (C) and one or more slave Nodes paired to the Coordinator. Both Coordinator and slave Nodes would be configured also as a gateway (GW) providing connectivity to other standards. Multi-bonding mechanism enables in this case the blue node N4 to work as a slave Node in the Network 1 and simultaneously create own Network 2 as its Coordinator. Listening communication in both networks, some packets received in the Network 1 would be forwarded to the Network 2 and vice verse. Specific behavior would be defined by application layer. This mechanism would be used for bridging networks by just few instructions of application code (Microrisc 2008b), would be used in telemetry in power sensitive applications to reduce number of

hops by collecting data from one networks and sending them together. It would used also as a arbitrage mechanism to avoid interfering of two or more networks: One Coordinator would coordinate Coordinators of the other networks, e.g., dedicate time slots to them. It is useful especially in one channel environment, e.g., wireless systems based on ASK (Amplitude Shift Keying) modulation.
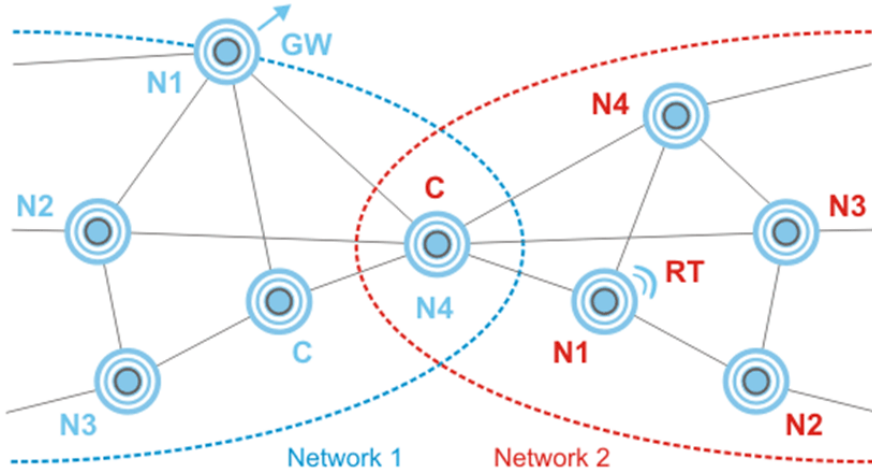

Fig. 5. IQMESH network chaining

Patented transceiver architecture having two layers (basic routines and application layer) provides an easy way to reduce development costs when creating connectivity applications. Transceiver modules already include protocol support in the Basic layer (would be referred also as a Operating System, Basic Routines, Protocol Layer, etc.), while the behavior of the device would be customized by Application layer utilizing routines from the Basic layer. In opposite to Solution stack, there is no need to compile protocol related routines, just application, consequently, it saves time of application development.

A special signal coding scheme brings higher data throughput due to real time data compression and also higher reliability and noise immunity due to perfect DC balance of the coded signal (Sulc 2007c).

Patented direct peripheral addressing in wireless networks provides an easy way to make open communication platforms utilizing built-in IQMESH features (Sulc 2008).

3.1 IQMESH basic

IQMESH uses mesh network topology supporting up to 240 devices: one Coordinator (C) mastering the network and up to 239 slave Nodes (N). It brings efficient addressing scheme using just one byte both for addressing of device and groups. Each Node provides background routing service for network packets or can be configured as a dedicated router (RT). Both Coordinator and Node can be setup as a Gateway (GW), specialized device bridging IQMESH network and other standards.

IQMESH protocol supports both individual and group addressing, as well as a network broadcasting. Besides standard features like bonding and discovery it also supports also direct peripheral addressing.

IQMESH protocol was defined as a light and portable to the inexpensive hardware with limited resources. Therefore, one byte internal addressing scheme was chosen, enabling to address 240 devices and up to 15 groups.

3.2 IQMESH packets

IQMESH protocol supports a packet oriented communication scheme, both point-to-point and more complex networking topologies (star, mesh). IQMESH protocol is flexible and leaves possibility for future expansion. For instance, there is a byte in the NTW INFO section of the packet defining routing algorithm. This simple mechanism allows to implement and support more rating algorithms and/or to have them application oriented as every application has usually very different requirements. For example, typical Smart House application would be realized with 4-hops and there is a need for fast response, while collecting data from power meters needs usually network supporting much more hops is needed, latency is not a problem.

	D	C		C		C	C
P I N	L E N	R C H	NTW INFO	R C N	DATA	R C D	R C S
PAH			NTWINFO		DATA		
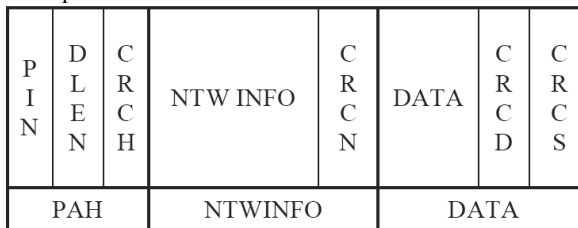
Fig. 6. IQMESH packet structure

Based on application layering, every device can accept and/or reject peer-to-peer communication. Packets for peer - to- peer communication consists of two block - PAH (packet header) and from DATA, while packets for networking communication consists of three blocks - PAH (packet header), NTWINFO (networking information) and DATA. Every block has its own consistency check mechanism to achieve high reliability even in a noisy environment. Basic packet structure is shown at Fig. 6.

PAH (packet header), 3 bytes long block, carries basic information about a packet, such as data length, flags if a packet is intended for peer-to-peer or networking communication, flags indicating system communication, flags indicating routing, direct peripheral addressing, such as encryption and acknowledgment request. NTWINFO (networking information) block has variable length based on PAH flag definitions. This mechanism provides an easy, reliable, while highly complex way to fit many different application needs. For example, Star topology does not need additional routing information which is requested in mesh networks. Setting ROUTEF = 0 will make a packet suitable for Star topology networks, while after setting ROUTEF = 1 six bytes describing the rating algorithm will be added to the NTWINFO.

Data load would vary between 0-255 bytes, while specific IQMESH implementations would support only 64bytes of data. This limitation enables porting of IQMESH protocol even to the smallest 8b microcontrollers.

Detailed IQMESH protocol description and its specifications will be publicly open and available in June 2009 (Microrisc 2009a).

4. Future work

In this chapter, only main functions of the IQRF platform are described. The basic modules are using 8bit microcontrollers (Microchip 2005) with limited space for the end user program. The newest modules are using 16bits microcontroller where is bigger place for user program, for implementation of some security aspect and so on.

IQMESH protocol is scalable allowing future expansion of routing algorithms. Currently new multi channel multihop algorithms utilizing all advantages and unique features of IQMESH protocol are under development.

The future step to simplify application development is to standardize peripherals and services sets (further on referred as HWP profile) provided by a specific application family. For example, a light switch would interpret data in a packet as I/O vector enabling R/W operation to the respective I/O pins of a transceiver module. In addition to the above I/O function, the transceiver module would support standard services like a bonding to and unbinding from the IQRF network. In this case, the application layer of the module will include a program sequence, interpreting packets as commands for R/W operations enabling access to peripherals and services of the module.

5. Conclusion

IQRF is new wireless communication platform for home/office and industrial automation. It has its own operating system for fast and easy implementation to user application.

The platform development tool contains software and hardware resources for rapid application development and prototyping.

IQRF platform implements IQMESH protocol. The protocol was defined as a light and portable to the inexpensive hardware with limited resources. Therefore one byte internal addressing scheme was chosen, enabling to address 240 devices and up to 15 groups. IQMESH protocol supports networks with up to 240 devices, one Coordinator and up to 239 slave Nodes. Each Node provides background routing service for network packets. Both Coordinator and Node can be setup as a Gateway (GW), specialized device bridging IQMESH network and other standards. IQMESH protocol can be fully or partially ported to the smallest 8 bit microcontrollers.

IQMESH implements several unique and patented features - a special signal coding scheme brings higher data throughput, higher reliability and noise immunity, two layer transceiver architecture reduces development costs, simultaneous functioning of devices in two or more wireless networks allows network chaining and finally, the mechanism of direct peripherals addressing in wireless networks directly supported by IQMESH protocol provils an efficient tool to build up open platform for wireless communication.

IQMESH protocol is scalable and ready to support new routing algorithms. All currently supported routing schemes are ported to the smallest 8b microcontrollers. IQMESH protocol definition will be opened, as well as public release of the definition, in June 2009. (Microrisc 2009a)

6. Acknowledgement

7. References

Aurel. (2008). "Radiomodem & Data transceivers." <http://www.aurelwireless.com/rf-wireless-modules/data-transceivers.asp> (May 8, 2009.

De Nardis, L., and Di Benedetto, M. G. (2007). "Overview of the IEEE 802.15.4/4a standards for low data rate wireless personal data networks." In: 4th Workshop on Positioning, Navigation and Communication 2007 (WPNC 07), T. Kaiser, K. Jobmann, and K. Kyamakya, eds., Hannover, GERMANY, 285-289.

Ferrari, G., Medagliani, P., Di Piazza, S., and Martalo, M. (2007). "Wireless sensor networks: Performance analysis in indoor scenarios." Eurasip Journal on Wireless Communications and Networking.

Flowers, D., and Yang, Y. (2008). "MiWi Wireless Networking Protocol Stack."

Ghazvini, M. H. F., Vahabi, M., Rasid, M. F. A., and Abdullah, R. (2008). "Improvement of MAC Performance for Wireless Sensor Networks." In: 13th International-Computer-Society-of-Iran-Computer Conference, H. Sarbazi-Azad, B. Parhami, S. G. Miremadi, and S. Hessabi, eds., Kish Isl, IRAN, 147-152.

Huang, Y. K., Hsiu, P. C., Chu, W. N., Hung, K. C., Pang, A. C., Kuo, T. W., Di, M., and Fang, H. W. (2008). "An Integrated Deployment Tool for ZigBee-based Wireless Sensor Networks." In: 5th International Conference on Embedded and Ubiquitous Computing, C. Z. Xu and M. Guo, eds., Shanghai, PEOPLES R CHINA, 309-315.

Chan, H. K. (2008). "Wireless Industrial Tracking System for Factory Automation." In: 2nd International Symposium on Intelligent Information Technology Application, Q. Zhou and J. Luo, eds., Shanghai, PEOPLES R CHINA, 862-866.

Ji, Z. Z., Li, Y., Lu, H., and Ieee. (2008). "The Implementation of Wireless Sensor Network node Based on ZigBee." In: 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, PEOPLES R CHINA, 3654-3657.

Leonard, J. (2007). "Non-Standard Solutions as Alternatives for Low-Cost Wireless Communications." In: Nikkei Electronics Asia.

Liang, L. L., Huang, L. F., Jiang, X. Y., Yao, Y., and Ieee. (2008). "Design and Implementation of Wireless Smart-home Sensor Network Based on ZigBee Protocol." In: International Conference on communications, Circuits and Systems, Xiamen City, PEOPLES R CHINA, 487-491.

Microchip. (2005). "PIC16F87/88 Datasheet." <http://www.microchip.com> (25.4.2009, 2009).

Microrisc. (2008a). "IQRF Transceiver Module Simple Block Diagram." b. p. tr 21a scheme, ed., Microrisc s. r. o.

Microrisc. (2008b). "TR-xxx-21A Transceiver Module Data Sheet." 6.

Microrisc. (2009a). "Detailed IQMESH protocol description and its specifications will be publicly open and available in June 2009." <http://www.iqmesh.org> (May 10, 2009.

Microrisc. (2009b). "Microrisc Web Page." <http://www.microrisc.cz/new/weben/index.php> (May 19, 2009.

RFM. (2009). "TRC101 300-1000 MHz Transceiver." <http://www.rfm.com/products/data/trc101.pdf> (May 19, 2009.

Song, T. W., and Yang, C. S. (Year). "A Connectivity Improving Mechanism for ZigBee Wireless Sensor Networks." 5th International Conference on Embedded and Ubiquitous Computing, Shanghai, PEOPLES R CHINA, 495-500.

Sulc, V. (2007a). "Czech Republic Patent PUV 16181 - Electronic transceiver module for network wireless communication in electric or electronic devices or systems." Microrisc s.r.o.

Sulc, V. (2007b). "Czech Republic Patent PUV 18340 - Module for wireless communication between electric or electronic equipment or systems, method for its control and method for creating generic platforms for user applications in area of wireless communications with those modules." Microrisc s.r.o.

Sulc, V. (2007c). "US Patent 7167111 - Method of coding and/or decoding binary data for wireless transmission, particularly for radio transmitted data, and equipment for implementing this method."

Sulc, V. (2008). "Czech Republic Patent PUV 18679 - A method of accessing the peripherals of a communication device in a wireless network of those communication devices, a communication device to implement that method and a method of creating generic network communication platforms with communication devices." MICRORISC s. r. o., Czech Republic.

Vojacek, A. (2007). "Bezdrátová komunikace z RS-232/485 - modul RC1280HP."

Z-Wave. (2009). "Z-Wave Technology Documentation." <http://www.zen-sys.com/modules/Zensys/> (May 3, 2009.

ZigBee. (2009). "ZigBee Aliance Web Page." <http://www.zigbee.org/> (May 5, 2009.

Wireless in Future Automotive Applications

Volker Schuermann, Aurel Buda, Stefan Jonker,
Norman Palmhof and Joerg F. Wollert
Bochum University of Applied Sciences
Germany

1. Introduction

Wireless technology became a part of the everyday life of many humans. Practically everyone possesses a mobile phone and is mobile attainable over it. Mobile phones became our daily way companions. Thus, and with the in the meantime clearly increased efficiency of these devices a number of new application scenarios are possible. Thereby can be fallen back on the experiences from other areas of application, for example from the mobile phone game market, which brought a quantity of interesting concepts out. Mobile phones increased not only their performance; they also bring along clearly a number of efficient communication interfaces, everything in front Bluetooth.

Also for the automobile industry the integration of mobile devices into vehicles is in the future an interesting market, since here completely new business models can be implemented. The chapter presents the boundary conditions for this. In addition also a concept for the integration of mobile devices in the vehicle belongs to it. Apart from pure aspects of communication, also the development and distribution of mobile applications are more near regarded.

First this chapter gives an overview of in the automotive environment spread communication technologies and their areas of application, the margin is here from short range technologies with ranges from few meters to long range communication over several kilometers away. Whereupon an overview of the key technology Bluetooth follows, whereby the emphasis honor on the application-oriented parts of the specification and the Bluetooth profiles is. Afterwards the Java Micro Edition, for the development of mobile applications, is in the focus of the chapter; here is a special attention, on the communication APIs and security. To the conclusion of the chapter then possibilities of the vehicle integration are described in detail on the basis of an example.

2. Wireless technologies and their areas of application

Wireless communication already belongs to the state of the art in many areas of the automotive environment today. Much works thereby covered off and is not noticed by the user of the end product. First of all a general overview of the assigned wireless technologies and their areas of application is to be given. See also for this Fig. 1.
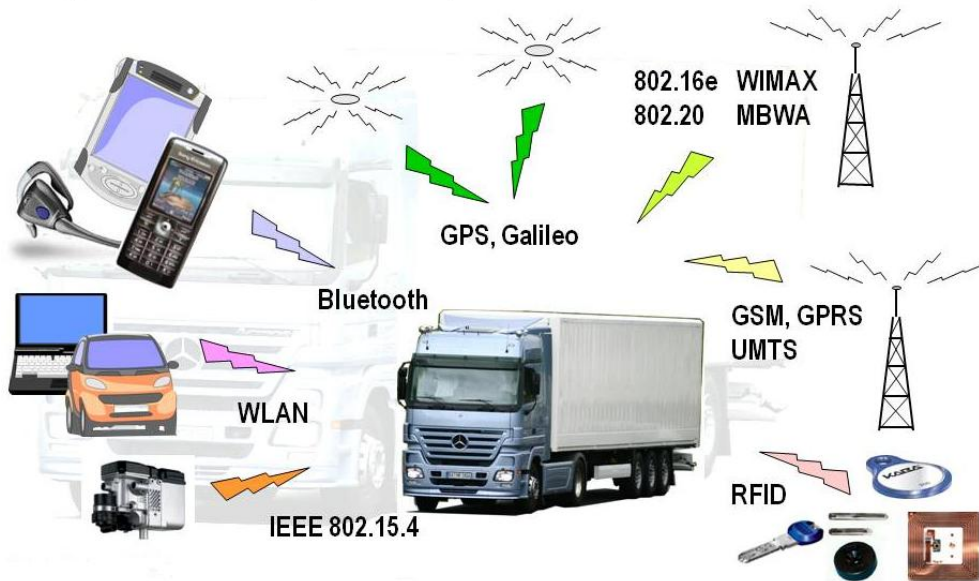

Fig. 1. Wireless Use Cases

IEEE 802.15.4:
IEEE 802.15.4 is a short range radio technology for wireless sensor networks. It forms the lower two protocol layers of a number of, in the automatic control engineering well-known, communication standards for example ZigBee or WirelessHART. The focus lies in the reliable transmission of small data sets if necessary over several hops away. In the automotive environment IEEE 802.15.4 is to be mainly found in production plants.

WLAN:
Main field of application of WLAN is the wireless integration of notebooks into local area networks. It looks similar also within the automobile area. Many of the diagnose tools necessary for modern vehicles, are today PC-based, which means a simple integration of WLAN, since both hardware and protocol stacks are present in large multiplicity at the market. So far these tools are usually connected over cables with the vehicle, whereby the diagnose unit must be in direct proximity to the vehicle. One is at present endeavored to replace these in many cases unpractical cable connections e.g. if the vehicle is on a lifting platform is, by wireless communication. First developments aim at the use of adapters, which are attached to the ODB2 (on board diagnosis) interface. A complete integration of WLAN in the vehicle is not impossible in the future. Further a set of comfort functions can be realized so, for example the transmission of vcard files from the email program of the PC

to the navigation system. In addition there are ambitions to use WLAN for Ad-hoc communication of vehicles among themselves and/or for communication of vehicles with their environment. One speaks in this connection of Car to Car and Car to Roadside communication.

Bluetooth:
Bluetooth is used in vehicles nowadays mainly for the free speech mechanism and the integration of headsets. In addition especially from the comfort and multimedia area a number of further meaningful applications can be realized, for example the playback of, on the mobile phone stored music files on the cars audio system. More to the application possibilities of Bluetooth follows in the further process of the chapter.

GPS, Galileo:
In the today's time almost nobody drives an unknown distance without navigation system. But not only for the comfort of the drivers is the knowledge of an exact position of importance. In logistic processes for example it is important to know, where certain goods or vehicles are at the moment. In times of just in time production a special meaning comes to that. The management of large fleets would not be possible without actual and exact position information. The American GPS represents the state of the art here at present. The European Galileo system up-to-date still is in the planning phase.

WIMAX, MBWA:
WIMAX (Worldwide Interoperability for Microwave Access) and MBWA (Mobile Broadband Wireless Access) are called colloquially frequently also „wireless DSL ".With them „the last mile" to customers is to be bridged to provide them with a DSL equivalent access to the Internet. Operational areas are in special infrastructure-weak regions. Both standards possess besides a mobile component, it permits the transfer of larger data sets over a distance of some kilometers to a moving vehicle.

GSM, GPRS, UMTS:
Beside pure telephony also data communication continues to move into the foreground with these technologies. A similar goal pursued as with WIMAX and MBWA, although with usually smaller data rates. However these technologies in many countries offer a surface covering net cover. In the remaining regions the net still is in the development.

RFID:
Especially in the luxury segment keyless entry and keyless go systems are a firm component of cars. The transponders necessary for it are frequently RFID chips characterized by a very small energy consumption which frequently get along even without battery, since they get the energy from the surrounding electrical field.

3. Bluetooth

The intention behind the development of Bluetooth (Bluetooth SIG, 2009) (IEEE, 2002) was replacing cables between individual devices such as mobile phones, PDA's, PC's, cordless mice, headsets etc. Important aspects thereby were on the one hand the costs of the

individual radio modules as well as the energy efficiency of the devices working usually on battery basis. Further it was enormously important to develop robust radio modules which are not damaged in case of transport of the mobile devices. The connection should remain unimpaired of other radio transmitters and be Ad-hoc, thus spontaneously to be developed. All these aspects considered until today with the advancement of the Bluetooth standard, additionally are the requirements to the transmission rate of such radio communications ever more largely. Bluetooth is in the meanwhile a world-wide accepted standard, which is very popular to due to its versatility and fail-safe characteristic also in the industrial area.

The Bluetooth architecture is essentially divided into three parts: the Bluetooth core, a protocol layer and an application profile layer. The Bluetooth core forms the IEEE 802.15.1 standard; it consists of the lower layers, which are necessary for communication. Fig. 2 shows the principle structure of the Bluetooth protocol stack. The subchapter begins with the Bluetooth core and its components, followed by some fundamental protocols, the different application profiles of Bluetooth are more near explained thereafter.
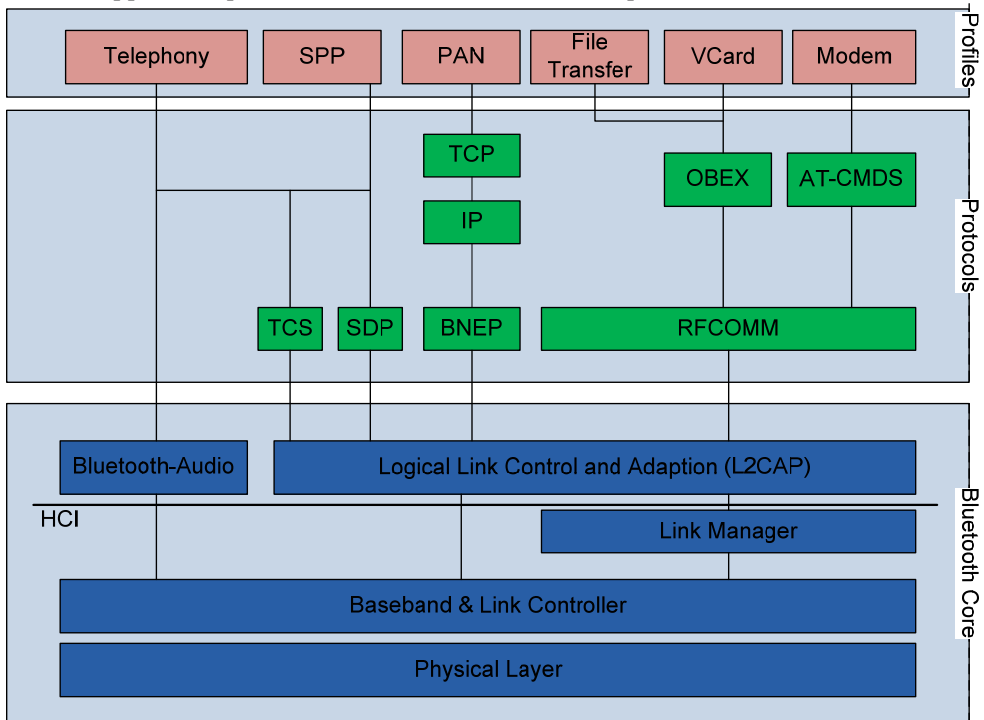
Fig. 2. Bluetooth Protocol Stack

3.1 Bluetooth Core

The Bluetooth core consists of several layers and forms the standard IEEE 802.15.1 in the actual sense. It covers the lower protocol layers beside the radio hardware also for the setting up of connections between devices. It is possible to set up piconets with up to eight active participants; one of them is the master of the piconet. Between the devices both

confirmed and unconfirmed communication for example for audio connections, can be established. For applications described in this chapter however the higher protocol layers and the services touching down on them are more interesting, therefore these are described in detail in the following.

3.2 Protocol Layer:

Above the Bluetooth core one finds a further layer, it consists of a multiplicity of different protocols, which represent the connection between Bluetooth core and application.

SDP:

In order to ensure the Ad-Hoc-ability of Bluetooth devices, is it necessarily that the devices between those a connection should be made, can recognize whether the other device supports the desired service. In order to manage this, the Bluetooth standard specifies the Service Discovery Protocol (SDP). Hereby it is possible to query the Service Record of a device. In the Service Records all available services of a Bluetooth device are stored, with a unique ID and service attributes. The inquiry of the Service Record is a Client-Server communication. The device, which would like to establish a connection to a service, sends an SDP Client inquiry to the SDP server of the other device, this sends in the response information about the supported services and it can be begun to establish a connection.

RFCOMM:

One usually used Bluetooth protocol is the RFCOMM (Radio Frequency Communications) - protocol. In principle RFCOMM is used everywhere, where a Bluetooth radio link should replace a physical cable, e.g. for the synchronization between a PDA and a PC. The RFCOMM protocol is able to administer up to 60 virtual serial interfaces at the same time. Other protocols like, the particularly in the mobile phone area spread, OBEX (Object Exchange Protocol) touches down on the RFCOMM protocol, a typical application of OBEX is the exchange of contact information between mobile phones or mobile phone and PC. Bluetooth replaces here with a radio link the device specific data cable. Likewise many Bluetooth profiles use the RFCOMM protocol, in the following with these is more in greater detail dealt.

TCS:

Telephony control Protocol specification (TCS) is the substantial protocol for the controlling of voice connections, all Telephony functions are regulated via this protocol.

BNEP:

The BNEP (Bluetooth Network Encapsulation Protocol) made possible like the name already says the encapsulation of different packets, which arise in a cable-bound network e.g. IP packets. Thus a Bluetooth device which is connected with the network over a Bluetooth Access point can exchange data and thus for example can use network printers. In order to realize this, the network packets are packed within BNEP frames, and passed to the lower protocol layers for transmission.

3.3 Bluetooth Profiles

The different Bluetooth profiles make the interaction of applications on different Bluetooth devices possible. They specify capability characteristics and parameters, which are needed, in order to communicate over certain protocols. They offer vertical access to the protocols. If two devices support the same profile, then they can communicate also spontaneously and problem-free with each another. Bluetooth is the only radio technology, which offers a so various service architecture. The most important fundamental profiles are presented in the following section. Fig. 3 shows the hierarchical layout of the Bluetooth profiles. An explanation of the abbreviations is effected in the following table.
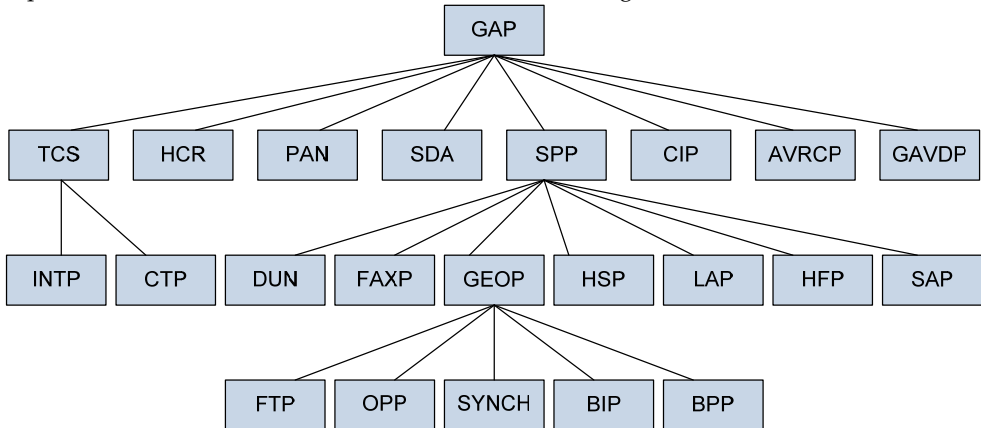
Fig. 3. Hierarchy of the Bluetooth of Profiles

GAP	Generic Access Profile
TCS	Telephony Control Specification Profile
HCR	Hardcopy Cable Replacement Profile
PAN	Personal Area Network Profile
SDA	Service Discovery Application Profile
SPP	Serial Port Profile Profile
CIP	Common ISDN Access Profile
AVRCP	Audio Video Remote Control Profile
GAVDP	Gerneric Audio Video Distribution Profile
INTP	Intercom Profile
CTP	Cordless Telephony Profile
DUN	Dial-up Networking Profile
FAXP	FAX Profile
GEOP	Generic Object Exchange Profile
HSP	Headset Profile
LAP	LAN Access Profile
HFP	Hands Free Profile
SAP	SIM Access Profile
FTP	File Transfer Profile
OPP	Object Push Profile
SYNCH	Synchronisation Profile
BIP	Basic Imaging Profile
BPP	Basic Printing Profile

Table 1. Bluetooth Profiles

GAP:

The Generic Access Profiles (GAP) is the most fundamental profile; here for all profiles fundamental characteristics are specified. To this the device name, the pin code and the Bluetooth address belong. Further functions are described such as connecting administration, operating mode and connecting security in this profile. GAP stands in the hierarchy of the Bluetooth profiles in highest place.

SDAP:

A further profile, which must be supported by all Bluetooth devices, is SDAP (Service Discovery Application Profile). It allows applications to access the already mentioned Service Record, which describes all services the device includes.

SPP:

The SPP is one of the usually-used profiles, because it offers the possibility of making up to 60 virtual serial interfaces available on a device. Each virtual interface possesses the characteristics of the well-known RS232 interface and reaches a data rate of 128kBit/s. The profile actually still serves as basis for further profiles. Fig. 4 describes which profiles SPP as basis for communication use.
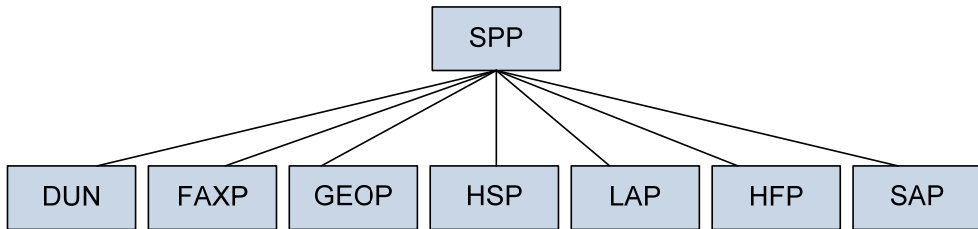
Fig. 4. SPP as basis

PAN:

The Bluetooth PAN profile was developed, in order to cover two concrete use cases. On the one hand it should be possible with the help of the PAN profile to develop a network infrastructure which equals wired LAN's. Further Access Point functionality with Bluetooth should be implemented, whereby the master functions as Access Point.

4. Java Micro Edition

The software for mobile devices is mostly written in C/C++ or Java. The question, which programming language is the best for the development of software on mobile phones could not be answered yet and is often the cause of discussions. But there are some serious reasons to use Java (Sun Microsystems, 2009) as programming language for mobile applications, because there is a Java virtual Machine for all important operating systems which runs the Java Byte Code on almost all devices without or with just very small changes. Thus a port from one system on to another one can be carried out with small expenditure. Restrictions with the port can occur with different hardware, like for example the minimal necessary display size. Java has a security API for security relevant operations like for example authentication and authorization. In the area of the mobile phones Java is very wide spread and has a large acceptance with all considerable hardware manufacturers. The Java Virtual Machine supervises the Java program, so that a crash does not affect other applications, thus the system becomes very robust. Further Java has, contrary to C++, a Garbage Collector which worries about the memory management. Due to nearly all important software enterprises and open source organizations cooperate in the Java Community Process (JCP), at the development and publishing of new specifications, there is a high market acceptance of Java. Because of these reasons Java has become a de facto standard for mobile application development and this will remain for a longer period.

Java Micro Edition (ME) (Breymann & Mosemann, 2008) is a subset of the Java standard Edition (SE) and is adapted on the needs of the mobile devices. Because the libraries of the standard edition are still much too large for the memory of the mobile devices. But in foreseeable time these devices will be efficient enough to use the full function range of the Java SE. Because of Moore's law the memory density on the chips doubles itself every 18 months.

Java ME is a collection of specifications and technologies, which particularly suit to the needs of the mobile devices. The structure of Java ME consists of configurations; these contain the Java Virtual Machine and a small sentence of class libraries from the Java SE. The configurations are extended by profiles, these contain further necessary APIs and optional Packages for special applications like for example Bluetooth communication. The

Java Micro Edition divides itself into two areas, depending upon features of the mobile devices.
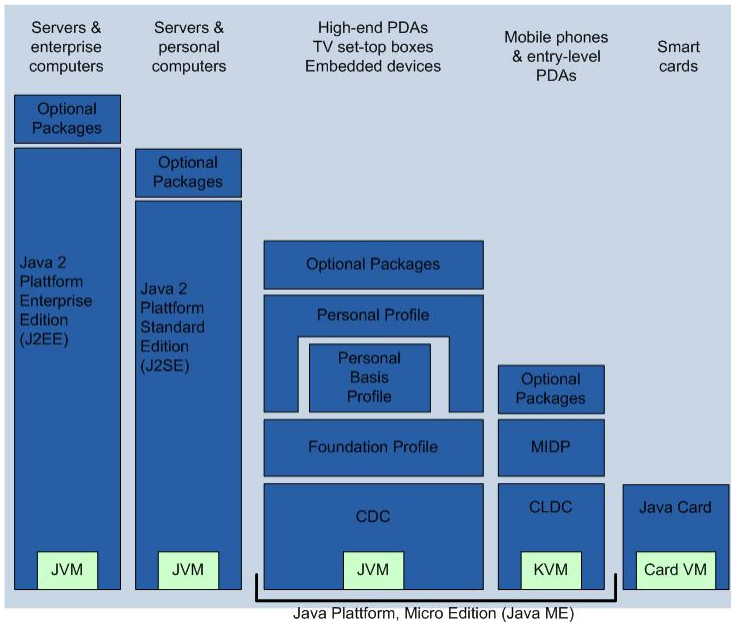

Fig. 5. Java ME in der Java Familie

For simple PDAs and mobile phones the Connected Limited Device Configuration (CLDC) is available. This contains the Kilobyte Virtual Machine (KVM). The configuration is extended by the Mobile Information Device Profile (MIDP). The MIDP builds up on the CLDC and extends this by a quantity of important functions, like for example the controlling of the life cycle of an application. For special applications the optional Packages are still available. The CLDC is designed for slow processors, little memory and unreliable network connections. The typically memory size is between 128 and 512 KByte. The CLDC and MIDP represent together a complete Java run time environment and are everything one needs to run simple Java programs on a mobile phone.

For more efficient devices like for example High end PDAs, set top boxes and embedded devices the Connected Device Configuration (CDC) is available, this contains a Standard Java Virtual Machine. To this run time environment belongs the Foundation Profile too, which makes the basic functions available for embedded systems. As well as the Personal Profile, this extends the Foundation Profile. In addition there are the optional Packages as with the CLDC. Fig. 5 shows the location of the Java Micro Edition in the surrounding field of the Java environment.
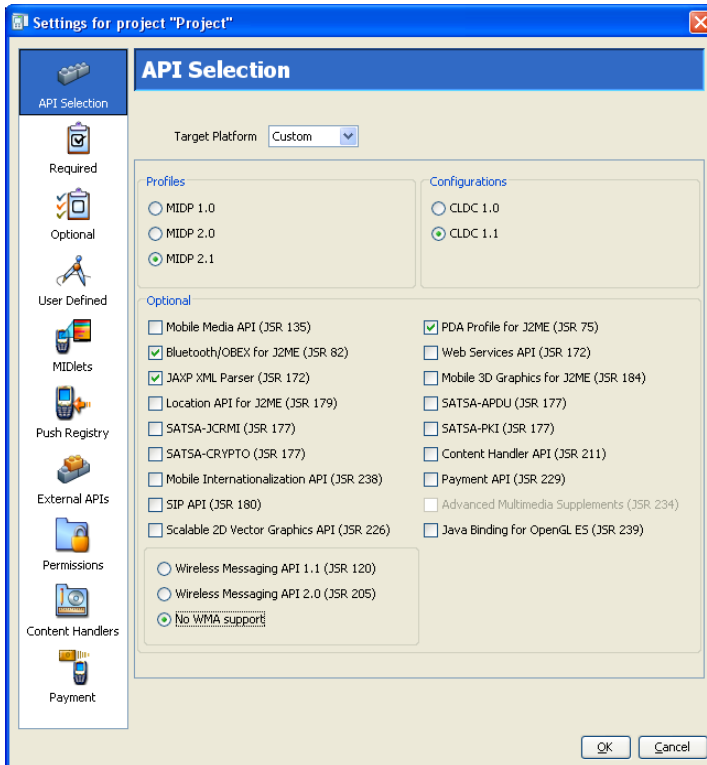
Fig. 6. Screenshot Wireless Toolkit: Selection of APIs

4.1 Configurations and Profiles

Since not all characteristics of all devices are known it is difficult to create a run time environment that fits to all characteristics of the different devices. Therefore one pursues the approach of configurations within the Java Micro Edition. A certain number of devices is assigned to a configuration according to their efficiency. Like that those programs are run able on all devices that have this configuration. The partitioning in CLDC and CDC configuration is made as in the previous section. Fig. 7 shows the architecture of a CLDC configuration with the MID Profile like it is used for mobile applications. Because of the rapid development and the short life cycle of such devices it is not possible to specify all variants. Therefore additional native applications, running direct on the operating system, and manufacturer-specific Java classes and applications are used.

	MIDP	Manufacturer-specific Applications	
Native Applications	Applications		Manufacturer-s. Classes
		MIDP	
	CLDC with KVM		
Operation System			
Hardware			
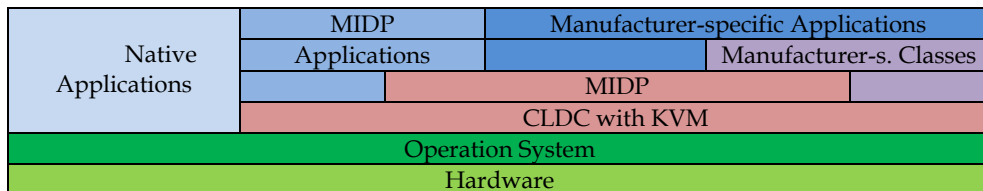
Fig. 7. Architecture of the CLDC

The following table shows an overview of at present available configurations of Java ME. Regarding to the technological development of the mobile devices with the CLDC 1.1 compared to its previous version 1.0 the minimum necessary memory was increased from 160 to 192 KB. The main reason for it was the introduction of the classes `Float` and `Double`. Further smaller errors were corrected and some additional classes were added. It might be only a question of time until all mobile devices support version 1.1 of the configuration, but at the moment one has to consider which version the current hardware supports.

JSR 30	CLDC 1.0	Connected Limited Device Configuration
JSR 139	CLDC 1.1	Connected Limited Device Configuration 1.1
JSR 36	CDC 1.0	Connected Device Configuration 1.0
JSR 218	CDC 1.1	Connected Device Configuration 1.1

Table 2. Java ME Configurations

The CDC configuration contains a substantially larger part of the Standard Edition APIs. And the appropriate Foundation Profile contains the entire Java Abstract Window Toolkit (AWT) with all functions necessary for executing Java Applets. The Personal Basis Profile is a subset of the Personal Profile and makes available nuclear functionality with a minimum graphic support. Here is not dealt with CDC and their profiles, since for the for the mobile application development the CLDC is crucial.

JSR 37	MIDP 1.0	Mobile Information Device Profile
JSR 118	MIDP 2.0	Mobile Information Device Profile 2.0
JSR 75	PDA	PDA Profile
JSR 46	FP	Foundation Profile
JSR 129	PBP	Personal Basis Profile
JSR 62	PP	Personal Profile

Table 3. Java ME Profiles

The MIDP bases on the CLDC and contains many important functions like for example network connections and their protocols, generation of sounds and user interfaces such as screen or keyboard. The Mobile Information Device Profile specifies also a set of minimal requirements to the hardware like for example a screen resolution of 96x54 pixels. Today the version 2.0 is supported by most mobile devices; an overview of the available Java ME profiles gives the above table.

The optional packages can be merged depending upon the needs of the application and the hardware requirements. Following table shows an excerpt of the most important Packages

with their JSRs numbers. All JSRs can be looked up under www.jcp.org the homepage of the Java Community Process, under the direction of Sun.

JSR 75	PIM	PDA Optional Packages (PIM und Dateisystem)
JSR 82	BTAPI	Bluetooth APIs
JSR 120	WMA	Wireless Messaging API
JSR 135	MMAPI	Mobile Media API
JSR 172		Web Services
JSR 177	SATSA	Security and Trust Services API
JSR 179		Location API
JSR 180		SIP API
JSR 184		Mobile 3D Graphics API
JSR 205		Messaging
JSR 211		Content Handler
JSR 226		Vector Graphics
JSR 229		Payment
JSR 234		Multimedia Supplements
JSR 238		Internationalization

Table 4. Optional Packages

Following table gives an overview of the spreading specifications. The Mobile Service Architecture specification (MSA) JSR 248 refers like its predecessor JSR 185 to a large extent of already existing specifications. It eliminates ambiguity and gives supplementing data where it is necessary. A goal of these spreading specifications should be to prevent a splintering of the individual APIs and give the different hardware manufacturers a guideline for the smallest common denominator. A device which fulfills the MSA specification must at least fulfill the MSA Subset, which is a subset of the MSA with decreased function range. An overview of the function range and the pertinent JSRs of the MSA and MSA Subset specification give Fig. 8. A minimum requirement to the hardware of the devices is also defined by the MSA specification. At least 1024 KB of volatile memory, a screen size of at least 128x128 pixels with a depth of shade of 16 bits. A multiplicity of further requirements can be inferred from the documentation JSR 248.

JSR 68	Java ME	Java ME Plattform Specification
JSR 185	JTWI	Java Technology for Wireless Industry
JSR 248	MSA	Mobile Service Architectur

Table 5. Package Bundles

		MSA Subset	MSA
JSR 238	Internationalization		
JSR 234	MultimediaSupplements		
JSR 229	Payment		
JSR 211	Content Handler		
JSR 180	SIP		
JSR 179	Location		
JSR 177	Security & Trust		
JSR 172	Web Services		
JSR 226	Vector Graphics	MSA Subset	
JSR 205	Messaging		
JSR 184	3D Graphics		
JSR 135	Mobile Media		
JSR 82	Bluetooth		
JSR 75	File & PIM		
JSR 118	**MIDP 2.0**		
JSR 139	**CLDC 1.1**		

Fig. 8. MSA and MSA Subset

4.2 JSR 82

The JSR 82 (JCP, 2008) was initiated by the JCP, for the development of Bluetooth based applications of communications and consists of the Java APIs for Bluetooth Wireless Technology (JABWT). This JSR represents no implementation of the general Bluetooth specification, but represents a collection of APIs for the configuration and controlling of the Bluetooth hardware in mobile devices.

The following subsections give beside the requirements of such a device and the structure of API architecture, views into the necessary configuration of services and devices and the general operational sequence of Java ME based Bluetooth communication under consideration of all security aspects.

Requirements:

For the employment of the JSR 82 API on mobile devices at least 512 KB main memory are needed, as well as a complete implementation of the Java ME CLDC version 1.0. In addition the existing Bluetooth hardware must exhibit a qualification of the Bluetooth Qualification Program at least for the profiles GAP, SDAP and SPP. Further the SDP, RFCOMM and the L2CAP profiles must be supported and accessibility for the API of these protocol layers must exist.

The access on the lower hardware and protocol layers is administered of a so-called Bluetooth Control Centre (BCC). Therefore it is not a component of the API, and must be provided by the hardware environment.

If all requirements are fulfilled, the Bluetooth API offers the following features during the application development:

- Registration of services

- Inquiry search of Bluetooth hardware and services

- RFCOMM, L2CAP and OBEX connections between Bluetooth devices

- Transmission of data, excluded voice connections
- Administration and controlling of communication connections
- Security mechanisms for expiration of communication

Here it is pointed out that the presence of Bluetooth and Java on mobile devices does not guarantee the support of the JSR 82 API, since among other things the possibilities of a device configuration are reduced by the Java ME. However this applies only to a part of the mobile phones offered nowadays.

Structure of API architecture:
The JABWT APIs extends the MIDP 2.0 platform with Bluetooth and OBEX support and consists of two packages, the fundamental Bluetooth API `javax.bluetooth` and the OBEX API `javax.obex`. Both are dependent on the package javax.microedition.io, which belongs to the CLDC, and optionally applicable depending upon requirements of the application. Fig. 9 clarifies the position of the Bluetooth API within an CLDC MIDP environment.
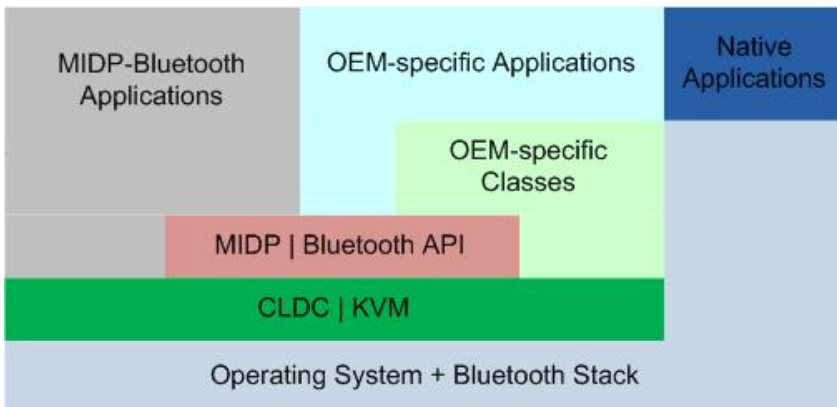
Fig. 9. Bluetooth in the Java Architecture

A Bluetooth application can be divided first into five ranges, which are processed with an implementation in chronological order: Stack initialization, management of devices, finding devices, finding services and communication. All APIs needed for these are part of the `javax.bluetooth` package.
As was already described on the basis the SDP, Bluetooth devices can take the role of a server or a client. This is specified in each case by the application. The activity diagram from following Fig. 10 gives an overview of the individual fields of server and client.
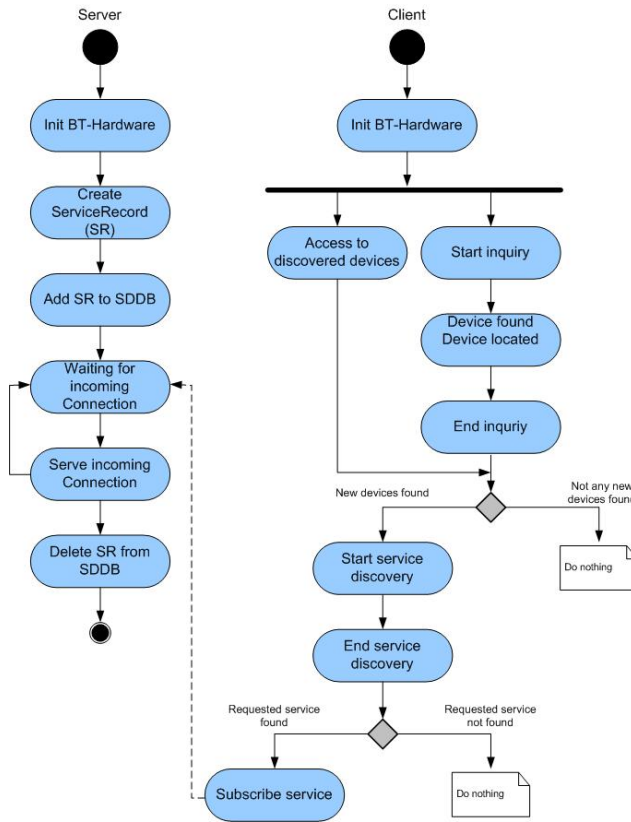
Fig. 10. Client and Server Activities

The initialization of the Bluetooth stack is independently of their operational area necessary for each Bluetooth application. A client application contains the search for devices and services, as well as the connection establishment with devices resulting from it and a following service use. A server application makes services available, administers these and reacts on connecting inquiries.

4.3 JSR 120 and JSR 205

A further Java API for the mobile communication is the Wireless Messaging API (WMA). The versions 1.0 and 1.1 were published in the JSR 120 (JCP, 2003) version 2.0 in the JSR 205 (JCP, 2004). With the Wireless Messaging API a mobile application can react on SMS and MMS messages, which are addressed to a certain port of the mobile phone, to which the application has registered itself, and process the received data. Messages also SMS in a binary format can be processed beside simple text or multimedia messages.

For further data communication in mobile communication networks as for example GPRS or UMTS further APIs are not necessary, since it concerns packet-oriented networks here and

so each mobile phone is IP addressable. The operating system usually makes this connection and administers it. From application view the standard APIs for Socket or HTTP communication can be used. It is the same procedure like in WLAN networks.

4.4 MIDlet

A Java program which was written for the MID profile is called to MIDlet; one or more MIDlets can be combined in a MIDlet Suite. After compiling source code one has a jad and a jar file, which can be loaded on a mobile phone afterwards. Each device on which a MIDlet should be executed must provide an environment which guarantees execution and administration of MIDlets. This environment is called Application Management Software (AMS) and controls the life cycle of the MIDlets. A MIDlet can be like the well-known Java Applet also only in one of three states. Between the two states Paused and Active the MIDlet can change during its runtime. The state Destroyed is however final. The MIDlet can even change its states by the help of special methods, but must notify the AMS about it. The AMS can change the states of the MIDlets at any time. This can happen if the resources of the MIDlets are needed by other processes, for example in case of a incoming telephone call the AMS sets the MIDlet into state Paused and the necessary display is used for the telephone call.

The MIDlet object is generated by the AMS and is first in the state Paused see Fig. 11, thus still no resources are blocked. Afterwards the MIDlet is started by the AMS through a call of the method `startApp()`. Now the MIDlet is in state Active and all needed resources will be requested. From the state Active the MIDlet can change again into the state Paused through the AMS or by itself. If for example a telephone call arrives the AMS sets the MIDlet into state Paused, since it needs some resources like for example the display of the MIDlet. The MIDlet asks periodically with the method `resumeRequest()` if it is allowed to run again, in this case the AMS starts the MIDlet by means of the method `startApp()`.
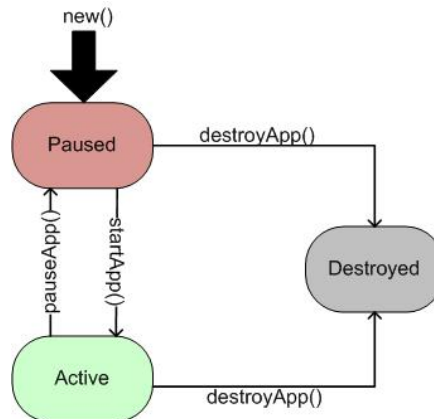
Fig. 11. State diagram of a MIDlet

From the state Active the MIDlet can be set by itself or by the AMS into state Destroyed. It releases then all requested resources and stores if necessary application data for the further use. Afterwards the MIDlet can be eliminated by the Garbage Collector.

4.5 Application Deployment

The occasionally complex installation was a big obstacle in the past which prevented a wide spreading of mobile applications. Usually for this a PC with for the mobile phone suitable configuration software was necessary, with which the mobile phone was connected by a data cable. For mobile Java applications there is another further alternative, which is favored in particular by the mobile games market. Here the installation of new MIDlets is at any time at each place within shortest time possible, always when the user needs certain programs for its mobile phone. This is reached by the download of the desired MIDlet over a UMTS or a GPRS connection. The necessary URL for this receives the user either from the browser of the mobile phone or by SMS. In addition such a call is also directly possible from a MIDlet. The development of specialized Part-MIDlets, for example for different equipment variants of a vehicle, is now possible which are downloaded on demand directly to the user's mobile phone.

The protocol for such a Over The Air (OTA) transmission is HTTP. Communication over HTTP is a firm component of MIDP and thus the standard technique for the data communication of MIDlets. The support of further protocols is however optional. In addition MIDlets offer with the method `platformRequest(string URL)` a standard procedure for the download of new programs over a HTTP connection.

Apart from the MIDP specification the optional content Handler API (JSR 211) contains also this functionality. Duty of the content Handler API is actually to pass certain tasks to other programs. For example playing music at the on the mobile phone installed media player. However it can be also used to download and to install new programs on the device.

With this kind of installation the appropriate jad and jar file must be on a web server reachable for the mobile device. In the jad file thereby to the location of the jar file is referred.

4.6 Security

In MIDP there is an extensive security concept, which on the public key procedure for the verification and authentication of MIDlet Suites is based. This security concept serves the preventing of, the use of sensitive operations, like for example the establishment of a expensive network connection, without preventing the knowledge of the user. So that a signed MIDlet can get access to a sensitive API, the appropriate permission must be set. This permission is indicated in the jad file.

In MIDP there are so-called Protection Domains which MIDlets are assigned to. In the Protection Domains is specified how to deal with the permissions.

There are the following Protection Domains:

- minimum: MIDlets of these Protection Domain, access to all Permissions is refused.
- untrusted: The user must give his agreement with each call to an API proteceted by a Permission of these Protection Domain. This is the default domain for unsigned MIDlets.
- trusted/maximum: The access to all Permissions of this Protection Domain is permitted.

One frequently still differentiates with trusted Protection Domains according to the certification authority:

• manufacturer: Uses certificates of the device manufacturer.

• operator: Uses certificates of the network provider.

• trusted third party: Uses third party certificates.

With the permissions two types are differentiated:

• allowed: The access is permitted without demand of the user.

• user: The user must give his agreement for the call of the associated API.

With user Permissions between the following types one differentiates:

• oneshot: Inquire with each call.

• session: Once inquire, decision remains valid as long as MIDlets of these MIDlet Suite are active.

• blanket: Once inquired, decision remains valid as long as the MIDlet Suite is installed.

If a MIDlet is in the trusted Protection Domain and the type of Permission is allowed, then it can use the associated API without demand of the user.

To which Protection Domain a MIDlet Suite belongs depends on the root certificate existing on the devices. With the installation the signature of the MIDlets is compared with the existing root certificates and accordingly a classification is made.

5. Vehicle integration

Cars are usually products, which come from one hand, from the car manufacturer. The offerers of accessory components so-called off board devices have a not insignificant problem, since usually no standard interfaces for the integration of these devices are present or must be licensed by the vehicle manufacturer. But even if such a license and the necessary installation interfaces are present, still the problem of the user interface remains for the offerer of accessory components. These are frequently goods in short supply and reserved for the OEM (Original Equipment Manufacturer) in the vehicle. From there the accessory offerers mostly offer their own control elements, which are expenditure-stuck or stuck on the instrument panel. Apart from the optical lack that control elements does not fit the design and cables lay partly openly, remains the problem, that these control elements do not fit into the control concept of the vehicle.

There is however one off board device, which is accepted by practically all car manufacturers and for both, interfaces for the integration in the vehicle and a firm place in the instrument panel is present. In addition it is suitable outstanding as universal control element for a multiplicity of devices. Meant here is the mobile phone.

Mobile phones are suitable on the one hand so well, because they possess many communication interfaces, beside the mandatory GSM, GPRS, UMTS support they frequently have Bluetooth and some models even WLAN interfaces. The employment of wireless technologies makes besides the cable to the control elements redundantly. The suitable communication technology can be selected depending upon application. For vehicle-internal communication a short range technology is sufficient as for example Bluetooth. However even if a genuine remote maintenance is to be realized over far distances a UMTS or a GPRS connection offers itself for this.

On the other hand mobile phones can be programmed almost at will, so that control applications for the most diverse devices can be realized. The advantages of the Java Micro edition in this area were stated already in detail.

5.1 Example auxiliary heating

How the integration into a vehicle is in detail realized is to be described in the following by the example of a auxiliary heating. The auxiliary heating is installed in the vehicle and attached to the CAN (Controller Area Network) bus of the car, over which all controllers are interlaced and receive their instructions. The instructions come of one at the instrument panel fastened or into it inserted, control element which is likewise connected with the CAN bus. Instead of this control element or also as addition of it now a mobile phone is to be used.

In principle for this UMTS/GPRS and Bluetooth present themselves as communication technology. Bluetooth for communication within the car and UMTS/GPRS for the remote maintenance from the domestic living room. Since the integration is very similar in both cases and Bluetooth besides brings the standardized communication profiles with it, contains the following example for the sake of simplicity only to Bluetooth. Following Fig. 12 outlines the fundamental structure of such a system.
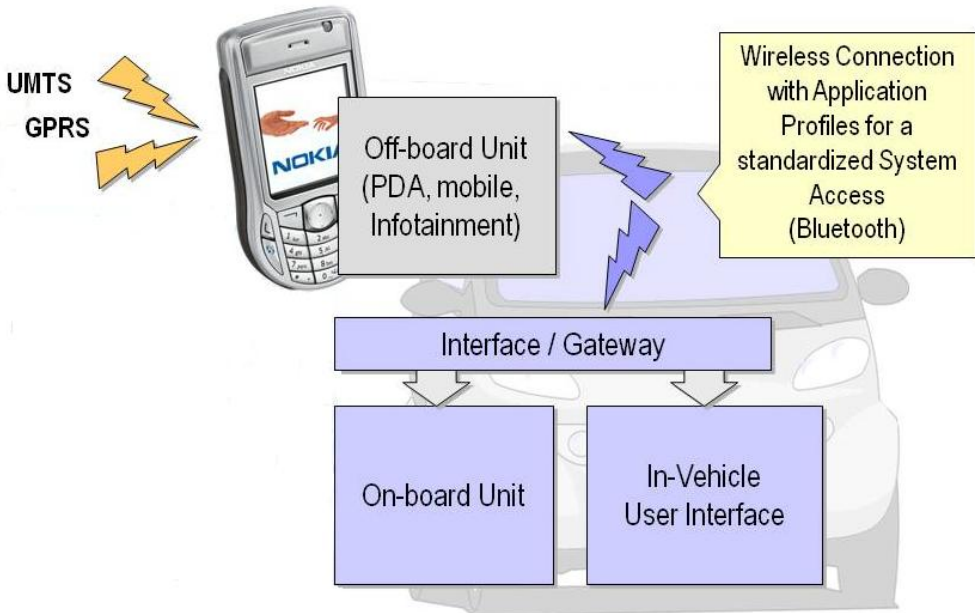
Fig. 12. Vehicle integration

The auxiliary heating and its control elements communicate no longer directly over CAN bus with each another, but over an interface or a gateway. The gateway controls the data transfer in the vehicle and passes the data on to the respective control devices. In the concrete example the gateway has a Bluetooth SPP connection to the mobile phone, over that it transfers the instructions of the remote control unit.

On the mobile phone a Java MIDlet runs, which the user downloaded ideally-proved directly from the Web server of the auxiliary heating manufacturer and installed it afterwards on his device. Security is ensured thereby by an appropriate signature of the MIDlets, which regalements the access to resources of the mobile phone e.g. communication interfaces and memory.

Even the selection of a suitable MIDlet for the vehicle-auxiliary-heating-mobile-phone-combination can be automated to a large extent, if device type and Bluetooth address of the user are deposited on a central server. This is can be done by a service technician for example with the installation.

The scenario to the deployment of the application has the following in Fig. 13 described expiration.
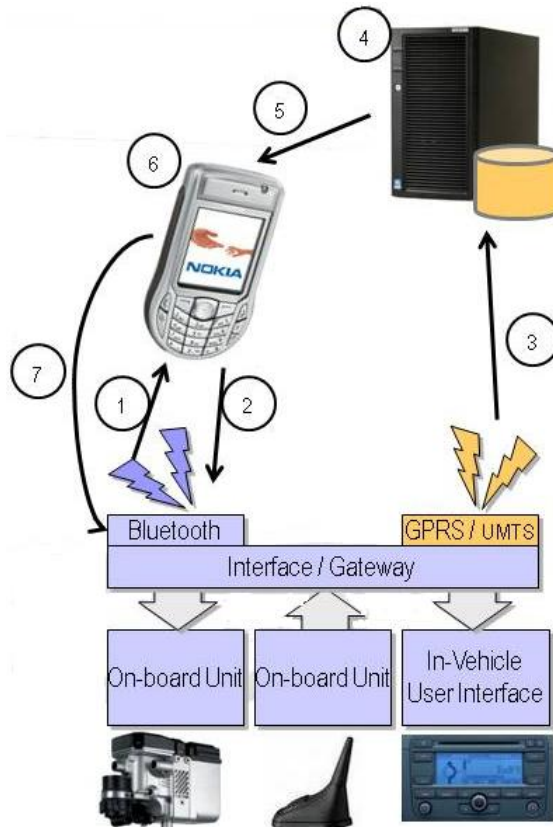
Fig. 13. Automatic Application Deployment

After the auxiliary heating was installed in the vehicle and the service technician has deposited the Bluetooth address, the telephone number and the type of device on the server starts the scenario.

1. The vehicle starts a search for Bluetooth devices in the environment. It acts around a functionality of the Bluetooth standard.
2. The found devices convey their Bluetooth address for later identification.
3. A list of the found devices is sent over a GPRS/UMTS connection to the server.
4. The server examined on the basis the Bluetooth addresses whether it for one of the found devices an order has, and selects on the basis the deposited type information a MIDlet fitting to the device type.
5. The server sends a SMS with the appropriate download link to mobile phone. Subsequently, the user opens the link and downloads the MIDlet to his mobile phone.
6. Subsequently, automatically the installation procedure begins. The user now only has to agree with the installation.

7. After finishing the installation the MIDlet is started directly and can be connected by Bluetooth with the auxiliary heating, in dependence of its signature. The application is ready for use thereby and the user can control his auxiliary heating.

6. Conclusion

The chapter showed that wireless communication already belongs in many areas of the automotive environment to the state of the art. Straight development possibilities further with the integration of mobile phones are present nevertheless. Bluetooth presents itself here as almost ideal communication technology for many applications.

For the development of mobile phone applications the Java Micro Edition is first choice, it offers not only large platform independence, but also detailed concepts to the deployment of applications or for security. In addition APIs are available for all usual communication technologies.

The mobile phone is the only off board device accepted by car manufacturers. That makes it interesting for the manufacturers of accessory components to use these as control elements. A concept for this was explained in the chapter.

7. References

Bluetooth SIG (2009). www.bluetooth.org

Breymann, U. & Mosemann H. (2008). *Java ME. Anwendungsentwicklung für Handys, PDA und Co (Germann)*, Hanser, 3446229973, Munich

JCP (2008). http://jcp.org/en/jsr/detail?id=82

JCP (2003). http://jcp.org/en/jsr/detail?id=120

JCP (2004). http://jcp.org/en/jsr/detail?id=205

IEEE (2002). http://ieee802.org/15/pub/TG1.html

Sun Microsystems (2009). http://java.sun.com/

Passive Wireless Devices Using Extremely Low to High Frequency Load Modulation

Hubert Zangl, Michael J. Moser, Thomas Bretterklieber, and Anton Fuchs
Institute of Electrical Measurement and Measurement Signal Processing,
Graz University of Technology, Kronesgasse 5, A-8010 Graz
Austria

1. Introduction

Whereas passive wireless communication in the Ultra High Frequency (UHF) domain features long ranges of several meters in free space, systems utilizing lower frequencies in the ELF (Extremely Low Frequency) to HF (High Frequency) domain can be advantageous in environments with conductive materials or where large antennas are not prohibitive. Additionally, the operation range is well defined and can be practically restricted to several centimeters like in Near Field Communication (NFC) *Standard ECMA-340 Near Field Communication Interface and Protocol (NFCIP-2)* (2003), although this does not necessarily mean that communication is secure Hancke (2008).

In this chapter we investigate passive wireless devices in the frequency range from almost DC to tens of Megahertz, i.e. from the ELF to the HF domain. Common abbreviations for the ITU frequency ranges are summarized in Table 1. The most common Radio Frequency Identification (RFID) systems use the LF (@125 kHz) and the HF (@13.56 MHz) bands. This chapter also considers lower frequencies.

Abbreviation	Range	Name	
subHz	< 3 Hz	SubHertz	
ELF	3 Hz - 30 Hz	Extremely Low Frequency	This Chapter
SLF	20 Hz - 300 Hz	Super Low Frequency	
ULF	0.3 kHz-3 kHz	Ultra Low Frequency	
VLF	3 kHz - 30 kHz	Very Low Frequency	
LF	30 kHz - 300 kHz	Low Frequency	
MF	300 kHz - 3 MHz	Medium Frequency	
HF	3 MHz - 30 MHz	High Frequency	
VHF	30 MHz - 300 MHz	Very High Frequency	
UHF	300 MHz - 3 GHz	Ultra High Frequency	

Table 1. ITU frequency ranges and abbreviations.

We provide a brief introduction to the technology, performance estimations in terms of powering range with respect to permitted signal levels and human exposure issues, performance considerations in terms of data transmission range with respect to background and man-made noise, and analysis of the impact of conductive/dielectric materials in the vicinity of the passive wireless devices (transponders).

We provide an introduction to the concept of load modulation techniques for passive wireless communication. Usually, RFID systems in the low to high frequency range (LF to HF) are considered as loosely inductively coupled transformers.

The basic principle of such mainly inductively coupled systems is shown in Figure 1. A primary coil of the reader generates an alternating magnetic field and induces a voltage in the coil antenna of the wireless device. The primary coil is connected to a diplexer that carries out frequency separation between the power supply path and the data path. The alternating magnetic field may penetrate layers of air, liquids (e.g. water) or layers of stainless steel and other weak conductors and will then induce a voltage in the coil antenna of the wireless device. A tuning element (e.g. a capacitor) and a power harvesting and storage unit (mainly comprising a rectifier and a storage capacitor) are needed to power the electronic components. A demodulator can extract data sent from the "reader". The transponder itself can transmit data by means of load modulation. This is, e.g., performed by a logic-controlled switch that changes the load of the secondary coil. The control logic can read out a sensor (e.g. a change in resistance or capacitance of a temperature or pressure sensor) and transmit this data back to the reader.
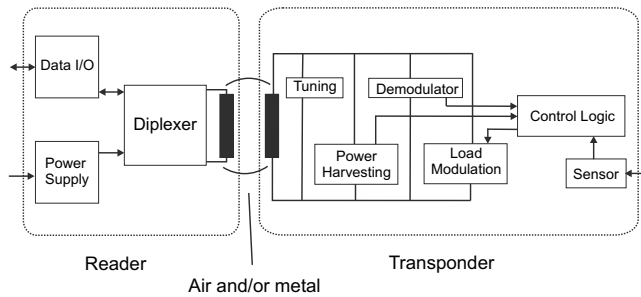
Fig. 1. Principle of passive wireless reader-transponder pair in the ELF to HF bands.

2. Application Examples

Various fields of application can be thought of for passive wireless devices in or behind metal housings powered by low frequent magnetic fields. In this section, we present example applications.

Transmitting measurement data through a double-walled stainless steel vessel can be necessary when extreme temperatures, high pressures, or other harsh environmental conditions (e.g. hazardous substances) are present. This could be a thermally insulated liquid hydrogen storage tank in a car, a whipped cream maker or a chemical reactor. Figure 2(a) shows a simplified block diagram of a setup. Figure 2(b) shows an example experimental setup using a whipped cream maker. The corresponding transponder is immersed in the liquid inside the vessel.

Another interesting application is e.g. for magnetic stirrers, which are standard devices in chemical laboratories. Equipping the magnetic stirring bar (a bar magnet with a protective coating made from either PTFE or stainless steel) with a miniaturized RFID tag that supplies a sensor interface, one could sense and display process parameters directly from the fluid that is stirred and thereby merge multiple devices (e.g. magnetic stirrer, temperature sensor, pH sensor) to a single device, which may need less space and reduce total equipment costs.
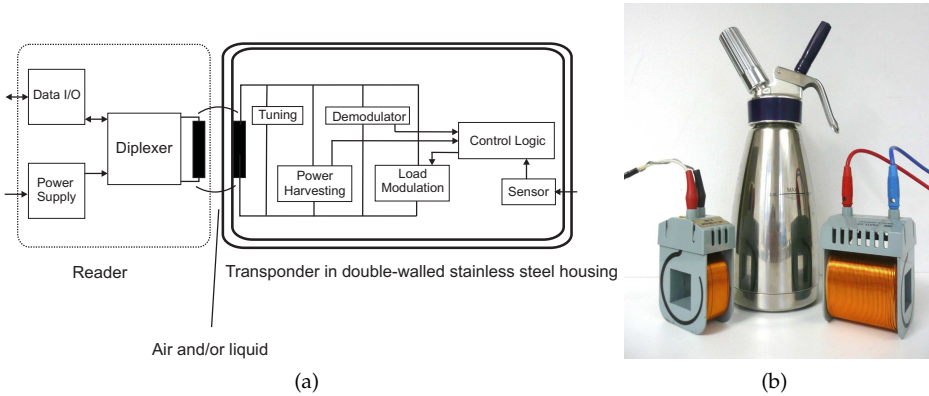
Fig. 2. (a) Schematic of a measurement setup for transmitting measurement data (e.g. pressure or temperature) through a double-walled stainless steel vessel. (b) Photo of an example measurement setup for transmitting temperature data (also other quantities such as pressure could be measured) through a thermally insulated double-walled stainless steel vessel. Here, the passive transponder is placed in a whipped cream maker and could be used, e.g., to monitor the temperature of the liquid.

Figure 3(a) depicts the schematic of such a setup, Figure 3(b) shows an example laboratory setup.
Figure 3(a) depicts the schematic of such a setup, Figure 3(b) shows an example laboratory setup

3. Comparison of Frequency Ranges from ELF to LF

The power that can be transmitted to wireless electronic devices by means of inductive coupling is rather limited by restrictions of the field strength than by technological limits. E.g., power transmissions of up to 60 W have been reported in Kurs et al. (2007). However, in practice we are faced with limitations of the permitted field strength, due to both electromagnetic compatibility and human exposure issues. The investigations in this section are based on the limits provided in *ERC Recommendation 70-03: Relating to the use of short range devices (SRD)* (2007) for limits regarding electromagnetic compatibility and ICNIRP (1998) regarding reference levels with respect to human exposure to alternating magnetic fields for the general public.

Electronic circuitry usually requires DC operating voltage. Therefore, passive devices require a rectifier circuit and an energy storage. Both diodes and transistors can be used for the rectifier, where transistors often offer the advantage of lower voltage drops. For the operation of the circuit it is important to achieve a certain minimum voltage. Consequently, the slew rate of the magnetic field must be high enough. This can be achieved by a high frequency and/or a high field magnitude.

Figure 4 shows Root Mean Square (RMS) reference levels for head, neck and trunk for the general public according to ICNIRP (1998). In the frequency range of up to 100 kHz, the peak values can be $\sqrt{2}$ higher than the RMS values. In the range from 100 kHz to 10 MHz, the permitted peak values increase to 32 times the RMS limit.
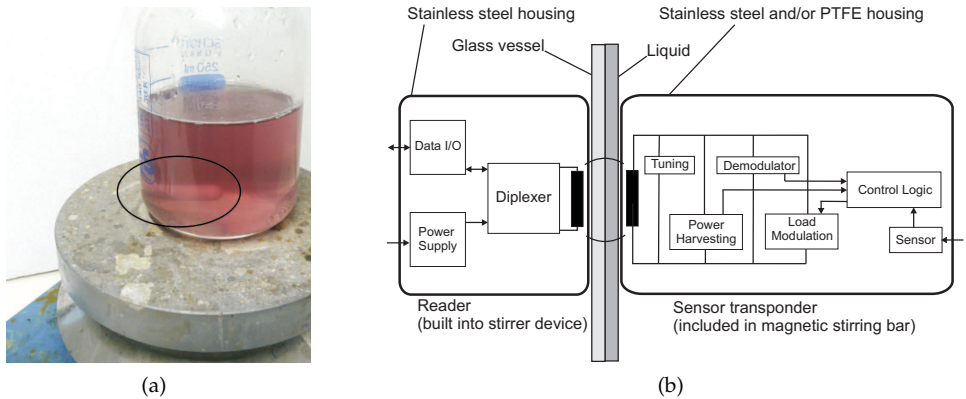
Fig. 3. (a) Photo of an example measurement setup for transmitting measurement data (e.g. temperature or pH data) from a sensor built into the magnetic stirring bar. The stirrer device will be equipped with a suitable readout circuitry and a numeric display. (b) Schematic of a measurement setup for transmitting measurement data.

An interpretation of these levels with respect to inductive wireless devices is provided in Figure 5. For a single turn circular loop with a square area of 1 cm^2, the voltage ranges from about 1 μV to several mV. However, this induced voltage is not sufficient to power an electronic circuit. With current semiconductor technology, the peak voltage should roughly exceed 1 V for a circuit to operate. Several techniques that can be used to increase the voltage are summarized in Table 2. The easiest methods are an increase of the area of the transponder antenna and an increase of the number of turns. Both methods are restricted by size and costs of the transponder. Resonance gain is also commonly exploited. Here, the antenna inductance L and an additional capacitor C form a resonance circuit. With this simple circuit, the antenna current and the voltage across the inductor as well as the capacitor are increased by the quality factor $Q = \frac{2\pi f L}{R}$ of the resonance circuit, which means that the coil resistance R must be low compared to the impedance of the coil inductance at the given frequency f. At lower frequencies, the resonance condition $f = \frac{1}{2\pi\sqrt{LC}}$ requires high L and/or C values, which may be difficult to implement. On the other hand, coils with high numbers of windings may have too low self resonance values due to parasitic capacitances e.g. in the HF range. Another drawback of high quality factors is the associated low bandwidth. A slight change of the inductance L or the capacitance C will change the resonance frequency of the circuit and the gain effect is lost. The resonance is also affected when two or more resonance circuits are in close vicinity. Therefore, in applications where many devices may be present (e.g. in batches of casino tokens) the quality factor is usually kept low. Further increases of the voltage can be achieved with electronic components such as diodes, e.g. in voltage multipliers (e.g. Gosset et al. (2008)) or in active up-conversion. The latter has the drawback that energy is required to get the up-conversion started (cf. section 3.2.1).

Usually, a combination of several of these techniques is necessary to make the low induced voltage useful for powering electronic devices. An example for the HF domain is provided in Figure 7. With several turns, an area of several square centimeters, and a quality factor above 10, the voltage can be sufficient to power the circuitry.
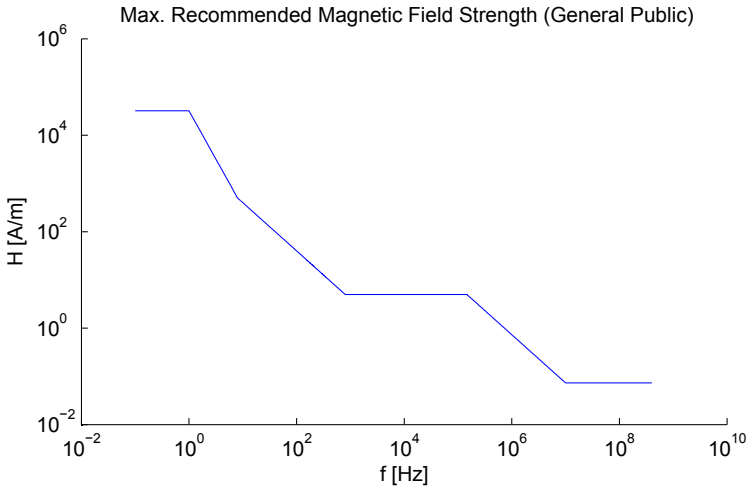
Fig. 4. Reference levels for the magnetic field strength for general public exposure to time varying fields ICNIRP (1998). These levels are obtained based on the impact (particulary on head, neck and trunk) of induced currents on the nervous system (up to 10 MHz) and the temperature increase of tissue due to absorption (above 100 kHz).
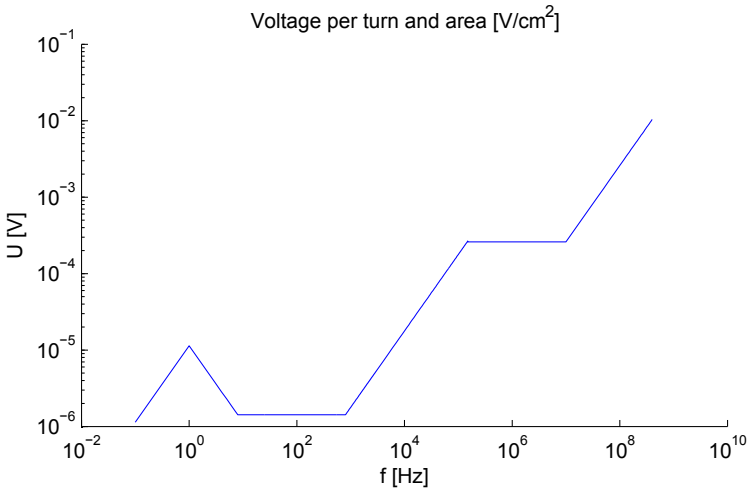

Fig. 5. Induced voltage for a single loop coil with an area of 1 cm^2 at the reference levels according to ICNIRP (1998).

Method	Disadvantages
Large number of turns	Low self resonance due to parasitic capacitance, costs
Larger area	Size, costs
Exploitation of resonance gain	Detuning, local increase of field strength, high C or high L for low frequencies
Ferromagnetic core (field concentrator)	Local increase of field strength, weight, costs
Voltage multiplier	Requires minimum induced voltage
Active conversion	Requires power up

Table 2. Comparison of methods for voltage enhancement.

With the small induced voltage at the reference levels for human exposure, e.g. in the ELF domain, one may wonder if this ranges can be of practical relevance. As long as it can be ensured that sensitive parts of humans will not reside permanently in the close vicinity of the reader devices, stronger fields can be used. In this case, it can be an advantage that the magnetic field strength decreases with the third power of the distance. However, besides limitation due to human exposure it is also mandatory that electromagnetic disturbances with respect to other devices are kept low. The permitted field strength is usually defined in a distance of 10 meters to the reader device. Therefore, it is possible for a certain antenna geometry to determine the maximum field strength at any distance in free air but also when the field is partially shielded, e.g. due to a metallic object. Limits according to *ERC Recommendation 70-03: Relating to the use of short range devices (SRD)* (2007) are shown in Figure 6. Based on these limits we can now determine the induced voltage at a certain distance.

Low frequencies offer the advantage that they are less affected by conductive material and have larger penetration depths. Consequently, such systems can be used for wireless sensing truly from the inside of, e.g., a steel object.

3.1 Environmental Influences

One of the major concerns for passive wireless communication is the reliability of the wireless link in the vicinity of conductive or strongly dielectric materials. In this section we will show that the use of low frequencies even permits communication through metal walls of e.g. several millimeters of stainless steel Zangl et al. (2008). Thus, a sensor can be placed inside of tanks without the need for cables or batteries.

The influence of a conductive wall on the magnetic field is illustrated in Figure 8 for a the range of 50 Hz to 50 kHz. Whereas the 50 Hz field is hardly affected by the wall, a significant attenuation occurs at higher frequencies. Therefore, lower frequencies are preferable for applications in the vicinity or through metallic objects. Recently, also an IEEE standard using low frequencies (131 kHz) in order to safely operate in the vicinity of conductive objects has been approved (*IEEE Standard 1902.1 for long wavelength wireless network protocol*, 2009). In this standard, also referred to as "RuBee", active communication rather than load modulation is used.

Often, the antenna inductance and a capacitor form a resonance circuit in order to increase the voltage in the transponder or the current in the reader. However, the resonance can be detuned when conductive or dielectric material is brought into the vicinity of the antenna. This has to be considered when a transponder is integrated into, e.g., wood or concrete. Otherwise the
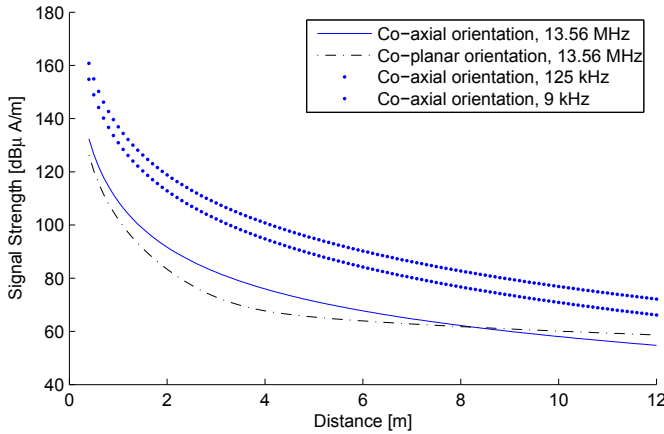
Fig. 6. Comparison of the permitted field strength according to *ERC Recommendation 70-03: Relating to the use of short range devices (SRD)* (2007) (based on a reader antenna of 20 cm times 30 cm). The graph can be used to determine the powering range. E.g., standard HF tags *Standard ISO/IEC 15693* (2006) are required to operate above 103.5 dBμA/m. Looking at the corresponding graph, this corresponds to a distance of about 1.6 meters. This could be slightly increased, e.g. by using a different antenna, but at this distance the shape has only minor influence. However, if a low power (low voltage) device can operate at about 80 dBμA/m (such as shown in Zangl and Bretterklieber (2007b)) the powering range extends to about 3 meters. For readers with lower field strength, the corresponding graphs just need to be shifted along the y-axis.
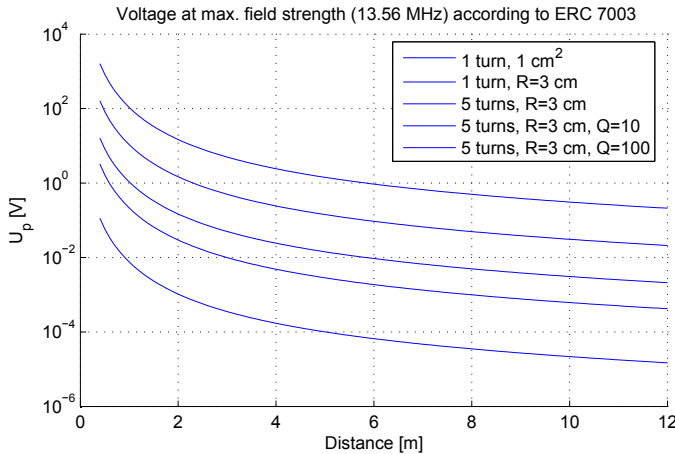
Fig. 7. Generation of the supply voltage: As the induced voltage per loop is very low, several techniques are used to increase the available voltage. Considering that current semiconductor technology starts to operate at about 1 V, a combination of the voltage enhancement techniques can yield sufficient voltage also at long distances to the reader.

performance will degrade. Antennas with low quality factors and non-resonant antennas are less sensitive to environmental conditions.
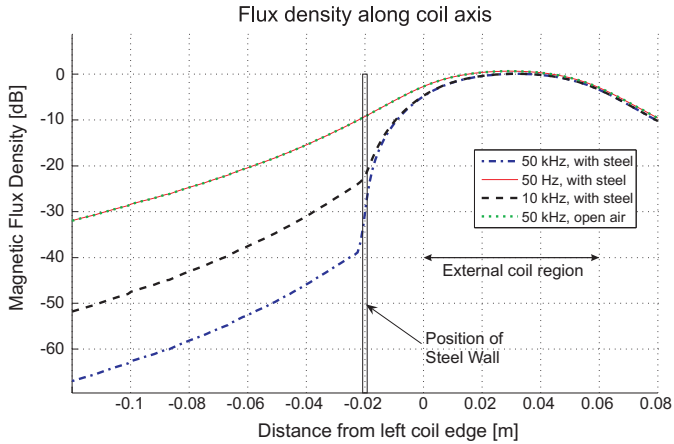
Fig. 8. Variation of the magnetic flux density along the rotational axis of the field coil for different frequencies, obtained by Finite Element Analysis. Regions of external coil and steel wall are marked by arrows. The curves for 50 Hz with steel and 50 kHz without steel (open air) coincide. Magnetic flux densities are referred to the maximum value, while the horizontal axis corresponds to the distance from the right edge of the steel wall. It can be seen that the steel wall hardly effects a 50 Hz signal while significant damping occurs at frequencies of 10 kHz and 50 kHz.

3.2 Data Transmission

With the ever decreasing power and voltage requirements of electronic components it can be expected that the powering range will further increase in the future. Does that mean that the operation range of passive wireless devices will also continue to increase? In situations where the powering range is the limiting factor, yes. However, with decreasing field strength and increasing data rates, another quantity becomes of major interest: The environmental noise. For a successful data transmission, the ratio E_b/N_0 between the bit energy E_b and the spectral noise density N_0 has to be sufficiently high; a reasonable value for good detection rates is a ratio of more than 10 dB Sklar (2001). Load modulation systems modulate the field that is generated by the reader. With increasing distance, the field to modulate becomes weaker and weaker and so does the generated response. Additionally, the distance for the data transmission also increases, which additionally lowers the E_b/N_0 ratio. The bit energy also decreases when we want to transmit at a higher data rate, as the signal to modulate and thus the total signal power remains the same, regardless of the bandwidth we use. Consequently, the modulated signal has to be as high as possible. Unfortunately, the environmental noise levels are very variable and are thus hard to predict; the sources in the considered frequency ranges are usually man-made. Reference levels can be found e.g. in *ERC Report 69: Propagation model and interference range calculation for inductive systems 10 kHz - 30 MHz* (1999).

When the amount of data is low, good E_b/N_0 ratios can also be obtained when energy is accumulated and then used for active transmission. On the other hand, when the coupling is
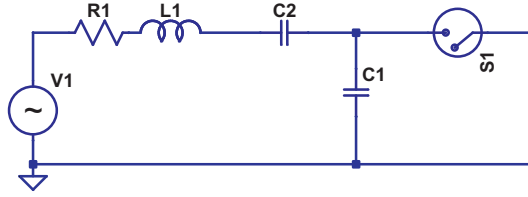
Fig. 9. Circuit for high data rate with high resonance gain. *V*1 represents the induced voltage due to the magnetic field generated by the reader; *L*1 represents the coil inductance and *R*1 the coil resistance. For a better illustration of the effect, the tuning capacitors *C*1 and *C*2 are of the same value. The tap point for the power supply rectifier circuit is between inductor *L*1 and *C*2, the switch *S*1 can, e.g., be an n-channel transistor. When the switch is closed, the resonance frequency is increased by a factor of $\sqrt{\frac{C1+C2}{C1}}$.

good and the modulation is strong, then high data rates (megabits per second) can be achieved Witschnig et al. (2007).

3.2.1 Resonant Modulation

Some of the techniques that we can use to increase the received voltage also increase the signal strength of the load modulated signal. This is true e.g. for larger area and resonance gain. Indeed, a quality factor $Q = 100$ means that the field strength at the position of the passive wireless devices is 100 times higher (with opposite orientation) when the device is present than when it is not. This may seem surprising as the device obtains energy from the field, but can be explained by the higher current compared to a short circuit loop. Furthermore, it could be assumed that a high quality factor would only permit low data rates, as the settling time for transitions between 0 and 1 would be proportional to the reciprocal of the bandwidth Finkenzeller (2003). However, as load modulation is a non-linear process, this is not necessarily always true as shown in Figures 10 and 11 for the circuit given in Figure 9. In this circuit the induced voltage is represented by *V*1. This model is valid when the reader is hardly affected by the transponder, i.e., when the coupling is low. For better coupling, the loading of the reader antenna has to be considered (Jiang et al., 2005).

3.2.2 Non-Resonant Modulation

A modulation circuit for non-resonant modulation, which is particularly useful at low carrier frequencies, is shown in Figure 12. Here, the modulation frequency is chosen higher than the carrier frequency. This has several advantages: Usually, up to HF the environmental noise level decreases with the frequency *ERC Report 69: Propagation model and interference range calculation for inductive systems 10 kHz - 30 MHz* (1999). Furthermore, the induced voltage in the reader coil increases as the slope of the magnetic field is increased. Additionally, this principle acts as a voltage converter, which can boost the voltage to a higher level.

A received signal at a receiver coil for a setup corresponding to Figure 2(b) is provided in Figure 15.

4. Reader Circuitry

A "reader" usually comprises the following components:

- Low noise power amplifier
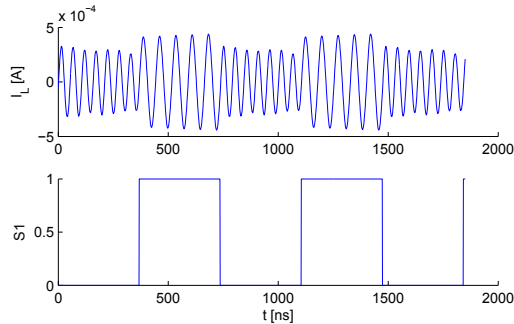
Fig. 10. Time signal for load modulation with resonance gain and high data rate. Ideally, the switching point for $S1$ is at the zero crossing of the voltage across $C1$, then no energy is lost. Provided that $C1 >> C2$ the energy loss also remains low even when the optimum switching point is missed. While the switch is closed, the resonance circuit continues oscillation but at an increased frequency. Looking at the carrier frequency in the spectrum, this means that the carrier signal is "turned off" immediately after closing of S1, no settling time is required. Once the switch is opened again (ideally at zero crossing of the voltage across $C2$) the signal immediately returns to the original frequency.
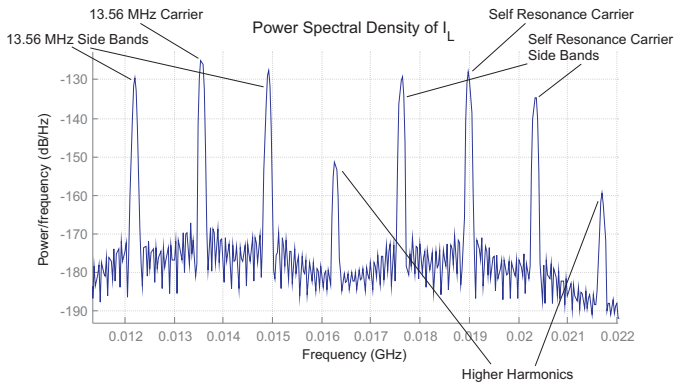
Fig. 11. Spectrum of the time signal of the coil current (proportional to the magnetic field) according to Figure 10. Besides the on-off amplitude modulation of the carrier, an alternating modulation of the switched resonance frequency with almost the same signal strength can be observed. With $C1$ being much larger than $C2$, the frequency difference would remain low and the spectra would overlap.
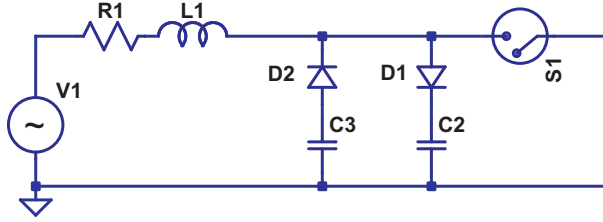
Fig. 12. Circuitry for non-resonant modulation with a low frequency for power transmission. $V1$ represents the induced voltage due to the magnetic field generated by the reader; $L1$ represents the coil inductance and $R1$ the coil resistance, $C2$ and $C3$ are energy storage capacitors.
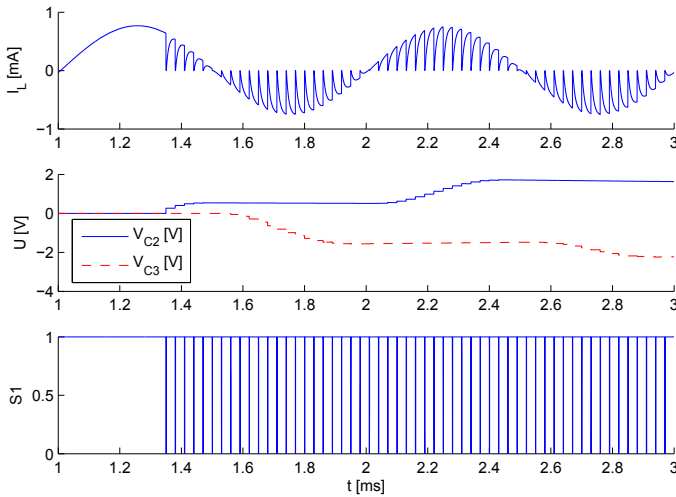
Fig. 13. Coil current and capacitor voltages for non-resonant modulation according to Figure 12. While the switch $S1$ is closed, the current in the coil increases due to the induced voltage $V1_p = 10$ mV, which is proportional to the rate of change of the field generated by the reader. When the switch is opened, the coil energy is rapidly transferred to capacitor $C1$ or $C2$ depending on the phase), such that the coil current returns to zero. Then, the switch $S1$ is closed again. The circuit does not only generate a modulation but also acts as a step up voltage converter.
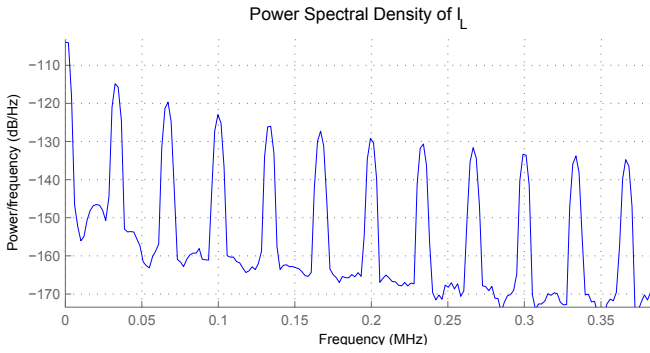
Fig. 14. Spectrum of the coil current according to Figure 13.
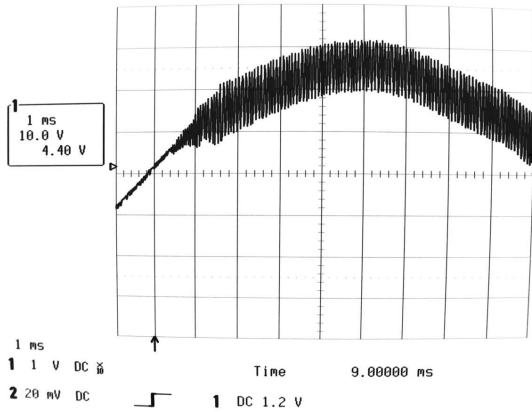
Fig. 15. Received voltage signal for non-resonant modulation, measured with a separate pick-up coil as shown in Figure 2(b). The modulation (higher frequency, here 1.6 kHz) generated by the passive wireless device and the signal from the reader (here 50 Hz) are superimposed but can be easily separated due to the large frequency offset. Even though no resonance is exploited, the received signal achieves a reasonable signal strength.

- Resonance loop antenna
- Carrier suppression
- Demodulation
- Symbol detection

Any of these components may be responsible for a certain limitation. For long range applications, high currents in the reader antennas are required. Often, resonance gain is also exploited for the reader such that the requirements for the power amplifier are eased. Linear and digital amplifiers are used, for the latter the resonance circuit also acts as a filter. In single antenna readers, the resonance loop acts as a filter for the signal received from the passive devices as well as for the noise, thus the E_b/N_0 ratio does not change. For high gain factors, the suppression of the modulated signal will be high such that the input referred noise of the receiver circuitry may dominate the environmental noise. In this case the E_b/N_0 ratio falls below the theoretical value and the performance degrades. In two antenna-readers, i.e. readers with a separate antenna for powering and receiving, this effect is not as important. Additionally, the pick-up antenna may be shaped (e.g. two opposite loops) such that the reader signal is suppressed, which eases demodulation and suppresses noise caused by the power amplifier. Other suppression techniques for the reader signal comprise active and passive filtering and directional couplers. Demodulation and detection of the signals are nowadays often performed in the digital domain. In this case the signal is sampled and processed on a Digital Signal Processor (DSP) or a dedicated hardware such as a Field Programmable Gate Array (FPGA). Such systems require an A/D conversion, which makes it mandatory to suppress the carrier. Demodulation can also be achieved with simple diode rectifiers in the analog domain. In this case, no additional carrier suppression is needed. Diode rectifiers can also be used to obtain inphase and quadrature (I and Q) signals (Zangl and Bretterklieber, 2007a).

5. Conclusion

The chapter presents passive wireless communication in the ELF to HF frequency range. With this technology, passive wireless devices can achieve ranges of up to several meters (at a low data rate), data rates of several megabit (at a low range). The devices can provide a well defined range of operation and they can permit communication in the vicinity or even through conductive or dielectric objects.

6. References

ERC Recommendation 70-03: Relating to the use of short range devices (SRD) (2007). *Technical report*.

ERC Report 69: Propagation model and interference range calculation for inductive systems 10 kHz - 30 MHz (1999). *Technical report*, European Radiocommunications Committee (ERC) within the European Conference of Postal and Telecommunications Administrations (CEPT), Marabella.

Finkenzeller, K. (2003). *RFID Handbook: Radio Frequency Identification Fundamentals and Applications*, 2nd edn, John Wiley & Sons, New York.

Gosset, G., Rue, B. and Flandre, D. (2008). Very high efficiency 13.56 MHz RFID input stage voltage multipliers based on ultra low power MOS diodes, *Proc. IEEE International Conference on RFID*, pp. 134–140.

Hancke, G. (2008). Eavesdropping Attacks on High-Frequency RFID Tokens, *Conference on RFID Security*, Budapest, Hungary.

ICNIRP (1998). International commission on non-ionizing radiation protection - guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz), *Health Physics* **74**(4): 494–522.

IEEE Standard 1902.1 for long wavelength wireless network protocol (2009).

Jiang, B., Smith, J., Philipose, M., Roy, S., Sundara-Rajan, K. and Mamishev, A. (2005). Energy scavenging for inductively coupled passive RFID systems, *Proceedings of the Instrumentation and Measurement Technology Conference 2005*, Ottawa, Canada, pp. 984–989.

Kurs, A., Karalis, A., Moffatt, R., Joannopoulos, J. D., Fisher, P. and Soljacic, M. (2007). Wireless power transfer via strongly coupled magnetic resonances, *Science* **317**(5834): 83–86.
URL: *http://dx.doi.org/10.1126/science.1143254*

Sklar, B. (2001). *Digital Communications*, Prentice Hall PTR, New Jersey.

Standard ECMA-340 Near Field Communication Interface and Protocol (NFCIP-2) (2003).

Standard ISO/IEC 15693 (2006). *ISO/IEC15693 Part 2: Air interface and initialization*.

Witschnig, H., Patauner, C., Maier, A., Leitgeb, E. and Rinner, D. (2007). High speed RFID lab-scaled prototype at the frequency of 13.56 MHz , *e&i Elektrotechnik und Informationstechnik* **124**(11): 376–383.

Zangl, H. and Bretterklieber, T. (2007a). Demodulation of 13.56 MHz load-modulated signals, *e&i Elektrotechnik und Informationstechnik* **124**(11): 364–368.

Zangl, H. and Bretterklieber, T. (2007b). Limitations of range of operation and data rate for 13.56 MHz load-modulation systems, *First International EURASIP Workshop on RFID Technology*, Vienna, Austria, pp. 7–12.

Zangl, H., Fuchs, A., Bretterklieber, T., Moser, M. and Holler, G. (2008). An investigation on wireless communication and power supply through metal tank walls, *IEEE International Instrumentation and Measurement Technology Conference*, 1452-1457, Vancouver Island, Canada.

UWB (Ultra wideband) wireless communications: UWB Printed Antenna Design

Abdallah Alshehri
Saudi Aramco
Saudi Arabia

1. Introduction

The admirable benefits of a wireless lifestyle have resulted in a huge demand for advanced wireless communications. The quick-tempered growth of the wireless communication market is expected to continue in the future since the claim of all wireless services is increasing. New wireless generations systems endeavor to provide high data rates as well as a wide range of applications like video and data to the portable users while supplying as many users as possible. However, this trend is limited by available resources like spectrum, power and coexistence of wireless devices. Thus, innovative technologies, that can coexist with devices operating on the crowded bands, are required to overcome the limited bandwidth and provide high data rates [Arslan, et al., 2006].

In February 2002, Federal Communications Commission (FCC) released a wide new unlicensed spectral band of 7.5 GHz for the commercial operation of ultra wideband (UWB) technology [FCC, 2002]. Since then, UWB has has been considered as one of the most promising wireless technologies to revolutionize high data rate transmission. UWB technology has offered unique advantages not achievable by conventional narrowband technology. These advantages are low power consumption, high speed transmission, immunity to multi-path propagation, and simple hardware configuration [Wentzloff, et al., 2005].

UWB communication techniques have attracted a great deal of interest both in the academia and industry in the past few years because of the high merit of their advantages. All Wireless systems and applications including UWB ones need a mean of transferring energy or signal from the apparatus to free space in the form of electromagnetic waves or vice versa which is an antenna. It has been recognized as a critical element of the successful design of any wireless device since the wireless systems are highly dependent on their antenna characteristics. Based on that, UWB antennas have been an important and active area of research and have presented antenna engineers with major design challenges [Alshehri, 2008].

In this chapter, the following topics are discussed. First, an introduction to UWB technology is presented in terms of its history, definition, advantages and applications. Second, the

importance of UWB antennas is highlighted. Next, the major requirements for a suitable UWB antenna are discussed. Then, several general methods to achieve a wide bandwidth are presented. After that, an overview on UWB antennas including UWB planar monopole antennas and UWB printed antennas is provided. Then, two novel designs of UWB printed antennas are introduced and investigated. Before we discuss these antenna designs in greater detail, we first introduce the numerical technique and its software package utilized to calculate the electromagnetic performance of the proposed antennas. The designs, optimizations, and simulations are conducted using the Ansoft High Frequency Structure Simulator (HFSS™). It works based on the Finite Element Method (FEM). After that, the design and fabrication of two novel UWB printed antennas are presented in details. The structural properties and performance characteristics of these antennas are investigated via numerical simulations and verified by measurements. The design process, parametric study, optimization as well as simulated and measured results, such as return loss, radiation characteristics and gain are provided.

1.1 UWB history

Now, Ultra Wideband technology is a potentially viable-revolutionary approach to wireless communication however it is certainly not a new concept. UWB systems have been historically based on impulse radio since it has transmitted very high data rates by sending pulses of energy instead of using a narrowband frequency carrier [Liang, 2006]. The concept of impulse radio dates back to the pulse-based spark-gap radio developed by Guglielmo Marconi in the late 1800's [Siwiak & McKeown, 2004]. It was used for several decades to transmit Morse code through the airwaves. But, it also caused strong interference to narrowband radio systems, which were developed in the early 1900's. Consequently, by 1924, the communications world abandoned wideband communication in preference of narrowband communication that was easy to regulate and coordinate [Schantz, 2003].

In the late 1960's, significant research was conducted by antenna designers, including Rumsey and Dyson [Rumsey, 1957: Dyson, 1959], who developed logarithmic spiral antennas, and Ross, who applied impulse measurement techniques to the design of wideband, radiating antenna elements [Ross, 1968]. As a result of these antenna advances, the development of short pulse radar and communications systems had begun. In 1973, the first UWB communications patent was awarded for the short-pulse receiver [Ross, 1973]. For the nearly 40 year period of 1960-1999, over 200 papers were published in accredited IEEE journals and more than 100 patents were filed on topics related to ultra wideband technology [Barrett, 2000].

On february 14th, 2002, FCC permitted the commercial operation of UWB technology [FCC, 2002]. After this official permission, research interest has exponentially grown with several researchers exploring RF, circuit, system and antenna designs related to UWB techonlogy. Also, several industrial companies have started investing in order to deliver revolutionary high-speed, short range data transfers and higher quality of services to the user [Powell, 2004].

1.2 UWB definition

Ultra-Wideband is a wireless communication technology that transmits an extremely low power signal over an extremely wide swath of radio spectrum to deliver very high transmission rates. It transmits and receives pulse-based waveforms compressed in time instead of sinusoidal waveforms compressed in frequency. Figure 1 depicts the equivalence of a narrowband pulse in the time domain to a signal of very wide bandwidth in the frequency domain. Also, it reveals the equivalence of a sinusoidal signal in time domain to a very narrow pulse in the frequency domain [Powell, 2004].
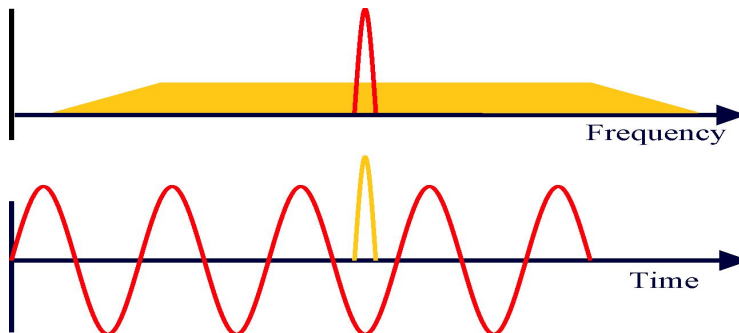
Fig. 1. The equivalence of pulse-based waveforms compressed in time to a signal of very wide bandwidth in the frequency domain [Powell, 2004]

In February 14th, 2002, The FCC allocated an unlicensed bandwidth from 3.1 GHz to 10.6 GHz to UWB applications, which was the largest spectrum for unlicensed use the FCC has ever granted. According to the FCC rulings, UWB operation is defined as any transmission scheme that occupies a fractional bandwidth greater than or equal to 0.2, or a signal bandwidth greater than or equal to 500 MHz [FCC, 2002].

1.3 UWB advantages

Due to the ultra wideband nature, UWB offers several motivating advantages that qualify it to be a more attractive solution to broadband wireless, radar communications and many applications than other technologies [Liang, 2006]. The extremely large bandwidth provided by UWB gives the potential of very high capacity resulting very high data rates. It can provide hundreds of Mbps or even several Gbps with distances of 1 to 10 meters [Oppermann, et al., 2004]. UWB communication systems transmit signals with extremely low spectral power density. By splitting the power of the signal across UWB spectrum, the effect is below the acceptable noise floor on any frequency. Consequently, UWB signals do not produce harmful interference to other coexisting wireless systems in the same frequency spectrum [Siriwongpairat, 2005].

Other merit of UWB includes low power transmission and robustness against eavesdropping since the UWB signal appears as noise-like signal. UWB also offers great flexibility of spectrum usage. In fact, this system can be designed as a function of the required data rate, range, power, quality-of-service, and user preference. By trading-off among these parameters, it can provide the required serivce based on the applications types

or provide service for multiple applications with a diversity of requirements devoid of additional hardwares [Arslan, et al., 2006].

1.4 UWB applications

As mentioned in the previous section, UWB offers many elegant advantages and benefits that are very attractive for a wide variety of applications. UWB is being targeted as a cable replacement technology since it has the potential for very high data rates using very low power at very limited range. It makes UWB became part of the wireless world, including wireless home networking, high-density use in business cores, wireless speakers, wireless USB, high-speed WPAN, wireless sensors networks, wireless telemetry, and telemedecine [Arslan, et al., 2006].

Due to the excellent time resolution and accurate ranging capability of UWB, it can be used in positioning and tracking applications such as vehicular radar systems for collision avoidance, guided parking, etc. The UWB capabilities of material penetration allows UWB to be used for radar imaging systems, including ground penetration radars, wall radar imaging, through-wall radar imaging, surveillance systems, and medical imaging [Oppermann, et al., 2004]. UWB radars can detect a person's breath beneath rubble or medical diagnostics where X-ray systems may be less desirable [Liang, 2006].

1.5 Why UWB antennas

The attractive nature of UWB coupled with the rapid growth in wireless communication systems has made UWB an outstanding candidate to replace the conventional and popular wireless technology in use today like Bluetooth and wireless LANs.

A lot of research has been conducted to develop UWB LNAs, mixers and entire front-ends but not the same amount of research has initially been done to develop UWB antennas. Later [Tsai & Wang, 2004; Lee, et al., 2004], academic and industrial communities have realized the tradeoffs between antenna design and transceiver complexity. In general, when new advanced wireless transmission techniques have been introduced, the transceiver complexity has increased. To maximize the performance of transceiver without changing its costly architecture, advanced antenna design should be used since the antenna is an integral part of the transceiver. Also, it has played a crucial role to increase the performance and decrease the complexity of the overall transceiver [Alshehri, 2004].

In addition, the trend in modern wireless communication systems, including UWB based systems, are to build on small, low-profile integrated circuits in order to be compatible with the portable electronic devices. Therefore, one of the critical issues in UWB system design is the size of the antenna for portable devices, because the size affects the gain and bandwidth greatly. The use of a planar design can miniaturize the volume of the UWB antennas by replacing three-dimensional radiators with their planar versions. Also, their two-dimensional (2D) geometry makes the fabrication relatively easy. As a result, the planar antenna can be printed on a PCB and thus integrated easily into RF circuits [Chen, et al., 2007].

2. Ultra Wideband Antenna Requirements

There are further challenges in designing a UWB antenna as compared to a narrowband one. A UWB antenna is different from other antennas in terms of its ultra wide frequency bandwidth. According to the FCC's definition, a suitable UWB antenna should provide an absolute bandwidth no less than 500 MHz or a fractional bandwidth of at least 0.2. This is the minimum bandwidth but generally the UWB antenna should operate over the entire 3.1-10.6 GHz frequency range resulting in spanning 7.5 GHz [Liang, 2006; Yang & Giannakis, 2004].

The UWB antenna performance is required to be consistent over the whole equipped band. Ideally, antenna radiation patterns, gains and impedance matching should be stable across the entire band [Wong, et al. 2005]. The radiation efficiency is another significant property of the UWB antenna. Since the transmit power spectral density is extremely low in UWB systems, high radiation efficiency is required because any unwarranted losses incurred by the antenna could affect the functionality of the system [Liang, 2006].

A suitable antenna should be physically compact and preferably planar to be compatible to the UWB unit, especially in mobile and portable devices. It is also greatly desired that the antenna attributes low profile and compatibility for integration with a printed circuit board (PCB) [Liang, 2006].

Finally, a UWB antenna should achieve good time domain characteristics. In narrowband systems, an antenna has mostly the same performance over the entire bandwidth and fundamental parameters, such as gain and return loss that have slight discrepancy across the operational band. Quite the opposite, UWB systems occupy huge operational bandwidth and often utilize very short pulses for data transmission. Consequently, the antenna has a more critical impact on the input signal. Indeed, minimum pulse distortion in the received waveform is a main concern of a suitable UWB antenna in order to provide a good signal to the system [Wong, et al. 2005].

3. Methods to Achieve Wide Bandwidth

As discussed in previous section, operating bandwidth is one of the most essential parameters of an antenna. It is also the main characteristic that distinguishes a UWB antenna from other antennas. Historically, a lot of effort has been made toward designing broadband antennas such as the helical antenna, biconical antenna and log periodic antenna. Most of these antennas are designed for carrier-based systems however their bandwidth is still considered narrowband in the UWB sense. Nevertheless, the design theory and experience associated with these antennas are very useful in designing UWB antennas [Lu, 2006]. Accordingly, several methods have been employed to widen the operating bandwidth for different types of antennas [Liang, 2006]. Some of these methods are explained in the following subsections.

3.1 The concept of frequency independence

The pattern radiation and the impedance characteristic of any antenna can be determined by its specific shape and size in terms of wavelength at a given operating frequency. However, a frequency independent antenna is an antenna that does not change its properties when its size has changed. This was first introduced by Victor Rumsey in the 1950's [Rumsey, 1957]. According to Rumsey's principle, the impedance and pattern properties of any antenna will be frequency independent if the antenna geometry is specified only in terms of angles irrespective of any particular dimensions. For this concept, there are basically three principles to achieve frequency independent characteristics. They are smoothing principle as in the biconical antenna, combining principle as in the log-periodic antenna and self-complementarity principle such as the case of spiral antenna [Alshehri, 2008].

3.2 The concept of overlapping resonances

In general, a resonant antenna has narrow bandwidth since it has only one resonance. However, the combination of two or more resonant parts, each one operating at its own resonance while living closely spaced together, may generate overlapping of multiple resonances resulting in multi-band or broadband performance. Actually, the two resonant parts technique has been broadly applied in antenna design, especially for mobile handset antennas that are required to operate at diverse wireless bands. The two resonant parts can be combined either in parallel [Chen & Chen, 2004], or one works as the passive radiator and the other as parasitic element [Muscat & Parini, 2001]. However, there is a main disadvantage of this concept. It can not provide constant radiation patterns over the operational bandwidth since the patterns differ from each other at different frequencies.

In theory, an ultra wide bandwidth can be attained by using a sufficient number of resonant parts provided that their resonances can be well-overlapped. Nevertheless, it is more difficult to practically obtain impedance matching over the entire bandwidth when there are more resonant parts. Furthermore, the antenna structure will be further complicated and expensive to fabricate. In addition, it is hard to have constant radiation characteristics when using multiple radiating elements [Liang, 2006].

3.3 The concept of increasing the radiator surface area

The conventional monopole is well-known antenna. It is composed of a straight wire perpendicular to a ground plane. It is one of the main antennas used widely in wireless communication systems due to its great advantages. These advantages include simple structure, low cost, omni-directional radiation patterns and ease for matching to 50Ω system [Balanis, 2005]. The -10dB return loss bandwidth of straight wire monopole is naturally around 10 %– 20 %, based on the radius-to-length ratio of the monopole [Liang, 2006].

The bandwidth of the monopole antenna increases with the increase of the radius-to-length ratio. This means that when the radius increases, the bandwidth will increase. In other words, the larger surface area (i.e. thicker monopole) will lead to a wider bandwidth due to the increase of the current area and thus the radiation resistance is increased [Rudge, et al. 1982]. Based on the concept of increasing the radiator surface area, instead of enlarging the radius of the conventional monopole, the wire is replaced with a planar plate yielding a planar

monopole. By using this technique, the bandwidth can be greatly enlarged. This planar plate can be designed using several shapes such as square, circle, triangle, trapezoid, Bishop's Hat and so on [Ammann & Chen, 2003; Agrawall, et al., 1998].

Many studies and analyses have been performed on the various shapes of the planar monopole antennas in order to understand their physical performance and to acquire enough knowledge of their operating principles. One study used the Theory of Characteristic Modes to determine how the planar monopole shape affects the input bandwidth performance of the antenna. Characteristic modes (Jn) are the real current modes on the surface of the antenna that depend on its shape and size but are independent of the feed point. These current modes produce a close and orthogonal set of functions that can be used to develop the total current. To characterize the electromagnetic behavior of electrically small and intermediate size antennas, only a few modes are needed, so the problem can be simplified by only considering two or three modes. This theory was used to analyze different planar monopole geometries such as square, reverse bow-tie, bow-tie and circular shapes. As a result of this analysis, the first characteristic mode J_1 was found to be similar to that of a traveling wave mode and its influence on the antenna impedance matching extends to high frequencies. Then, to obtain broad input bandwidth performance, it is necessary to obtain a well-matched traveling mode which can be achieved by reinforcing the vertical current distribution (mode J_1) and minimizing horizontal current distributions (mode J_2). This can be accomplished by using different techniques as will be discussed later [Bataller, et al. 2006].

A few simple formulas have been reported to predict the frequency corresponding to the lower edge of the -10 dB return loss impedance bandwidth for different shapes of the monopole antennas [Agrawall, et al., 1998; Evans, Amunann, 1999]. However, the prediction of the upper edge frequency requires full-wave analysis. Also, it is found that the upper edge frequency depends on the part of the planar element near to the ground plane and feed probe where the current density concentrates. Thus, different techniques are proposed to control the upper edge frequency such as beveling the square element on one or both sides of the feed probe [Ammann, 2001].

3.4 Techniques to improve the planar antenna bandwidth

Some shapes like the square and circular planar monopole antennas have a drawback of a relatively small impedance bandwidth [Ammann & Chen, 2004]. Consequently, several techniques have been suggested to improve the antenna bandwidth.

First, the radiator may be designed in different shapes. For instance, the radiators may have a bevel or smooth bottom or a pair of bevels to obtain good impedance matching. The optimization of the shape of the bottom portion of the antenna can lead to the well-matched traveling mode [Ammann & Chen, 2003].

Secondly, a different type of slot cut may be inserted in the radiators to improve the impedance matching, particularly at higher frequencies, [Chen, et al., 2003]. The effect of slots cut from the radiators is to vary the current distribution in the radiators in order to change current path and the impedance at the input point. Besides, using an asymmetrical

strip at the top of the radiator can decrease the height of the antenna and improve impedance matching [Cai, et al., 2005].

Thirdly, a partial ground plane and feed gap between the partial ground plane and the radiator may be used to enhance and control the impedance bandwidth. The feed gap method is crucial for obtaining wideband characteristics and it particularly affects mode J_1 (the vertical current distribution) resulting in the well-matched traveling mode [Agrawall, et al., 1998]. Also, a cutting slot in the ground plane beneath the microstrip line can be used to enhance the bandwidth [Huang & Hsia, 2005]. In addition, a notch cut from the radiator may be used to control impedance matching and to reduce the size of the radiator. The notch cut significantly affects the impedance matching, especially at lower frequencies. It also reduces the effect of the ground plane on the antenna performance [Chen, et al., 2007].

Fourthly, cutting two notches in the bottom portion of rectangular or square radiators can be used to further improve impedance bandwidth since they influence the coupling between the radiator and the ground plane. Also, transition steps may be used to enhance the bandwidth by attaining smooth impedance transition between the radiator and feeding line [Lee, et al., 2005].

Finally, several modified feeding structures may be used to enhance the bandwidth. By optimizing the location of the feed point, the antenna impedance bandwidth will be further broadened since the input impedance is varied with the location of the feed point [Ammann & Chen, 2004]. A shorting pin can be used to reduce the height of the antenna as used in a planar inverted L-shaped antenna [Lee,et al., 1999]. A double-feed structure highly enhances the bandwidth, especially at higher frequencies [Daviu, et al., 2003].

4. Overview on Ultra Wideband Antennas

Different kinds of wideband antennas are designed, each with its advantages and disadvantages. The history of wideband antennas dates back to those antennas designed by Oliver Lodge in 1897. Later, they led to some of the modern ultra-wideband antennas. These antennas were early versions of bow-tie and biconical antennas which had significant wideband properties. In the 1930's and 1940's, more types of wideband antennas were designed, such as spherical dipole conical and rectangular horn antennas. In the 1960's, other classes of wideband antennas were proposed such as wideband notch antennas, ellipsoid mono and dipole antennas, microstrip antennas and tapered slot and Vivaldi-type antennas. Also, frequency independent antennas were applied to wideband design like planar log-periodic slot antennas, bidirectional log-periodic antennas and log-periodic dipole arrays [Dotto, 2005].

The wideband characteristics of these antennas depend on two main antenna features, which are the geometry shape and the dielectric material type, if any. The antenna bandwidth is affected by the impedance match between the feeding circuit and free space. The bandwidth of these antennas fluctuates significantly, from hundreds of MHz to tens of GHz based on the antenna design [Dotto, 2005]. However, these antennas are rarely used in portable devices and are difficult to be integrated in microwave circuits because of their bulky size or

directional radiation. Alternatively, planar monopoles, dipoles or disc antennas have been introduced due to their wide bandwidths and small size. The earliest planar dipole is the Brown-Woodward bowtie antenna, which is a planar version of a conical antenna [Chen, et al., 2006].

4.1 Ultra wideband planar monopole antennas

Planar monopole antennas are constructed from a vertical radiating metallic plate over a ground plane fed by a coaxial probe. It can be formed in different shapes such as rectangular, triangular, circular or elliptical. The main features of these shapes are their simple geometry and construction. Planar monopole antennas have been explored numerically and experimentally and have shown to exhibit very wide bandwidth [Schantz, 2003; Ammann & Chen, 2003].

A circular monopole antenna yields a broader impedance bandwidth as compared to a rectangular monopole antenna with similar dimensions. This is because the circular planar monopole is more gradually bent away from the ground plane than the rectangular monopole. This provides smooth transition between the radiator and feed line resulting in a wider impedance bandwidth [Azenui, 2007].

The planar monopoles, suspended in space against ground plane, are not suitable for printed circuit board applications due to their vertical configuration. However, they can be well matched to the feeding line over a large frequency band (2 - 20 GHz) with gain of 4 - 6 dBi. But they suffer from radiation pattern degradation at higher operation frequencies [Chen, et al. 2006]. Therefore, some efforts have been made to develop the low-profile planar monopoles with desirable return loss performance in the 3.1 - 10.6 GHz frequency range. So, the antenna can be integrated to a PCB for use in UWB communications, which will be discussed in the following section.

4.2 Ultra wideband printed antennas

The UWB antennas printed on PCBs are further practical to implement. The antennas can be easily integrated into other RF circuits as well as embedded into UWB devices. Mainly, the printed antennas consist of the planar radiator and ground plane etched oppositely onto the dielectric substrate of the PCBs. In some configurations, the ground plane may be coplanar with the radiators. The radiators can be fed by a microstrip line and coaxial cable [Chen, et al. 2006].

In the past, one major limitation of the microstrip or PCB antenna was its narrow bandwidth characteristic. It was 15 % to 50 % of the center frequency. This limitation was successfully overcome and now microstrip antennas can attain wider matching impedance bandwidth by varying some parameters like increasing the size, height, volume or feeding and matching techniques [Bhartia, et al. 2000]. Also, to obtain a UWB characteristic, many bandwidth enhancement techniques have been suggested, as mentioned earlier.

Numerous microstrip UWB antenna designs were proposed. For instance, a patch antenna is designed as a rectangular radiator with two steps, a single slot on the patch, and a partial

ground plane etched on the opposite side of the dielectric substrate. It provides a bandwidth of 3.2 to 12 GHz and a quasi-omni-directional radiation pattern [Choi, et al. 2004]. A clover-shaped microstrip patch antenna is designed with the partial ground plane and coaxial probe feed. The measured bandwidth of the antenna is 8.25 GHz with gain of 3.20 - 4.00 dBi. Also, it provides a stable radiation pattern over the entire operational bandwidth [Choi, et al. 2006].

5. Ultra Wideband Printed Antennas Design

The planar antennas, printed on PCBs, are desired in UWB wireless communications systems and applications because of their low cost, light weight and ease of implementation. In addition, they can be easily integrated into other RF circuits as well as embedded into UWB devices such as mobile and portable devices. However, it is a well-known fact that the bandwidth of patch antennas is narrow. Thus, many attempts have been made to broaden the bandwidth of printed antennas.

Therefore, in this chapter, two novel designs of microstrip-fed printed antennas, using different bandwidth-enhancement techniques to satisfy UWB bandwidth, are introduced. According to their geometrical shapes, they can be classified into two types: the first type is a stepped-trapezoidal patch antenna. The second one is a double-beveled patch antenna.

In designing these antennas, it considers UWB frequency domain fundamentals and requirements, such as far field radiation pattern, bandwidth, and gain. The design parameters for achieving optimal operation of the antennas are also analyzed extensively in order to understand the antenna operation. It has been demonstrated numerically and experimentally that the proposed antennas are suitable for UWB communications and applications, such as wireless personal area networks (WPANs) applications.

Before we discuss these antenna designs in greater detail, we will first introduce the numerical technique and its software package utilized to calculate the electromagnetic performance of the proposed antennas. The designs, optimizations, and simulations are conducted using the Ansoft High Frequency Structure Simulator (HFSS™). It works based on the Finite Element Method (FEM).

5.1 Finite elements method (FEM)

The finite element method (FEM) is created from the need to analyze and solve complex structure analysis. The FEM is a partial differential equation (PDE) based method. FEM is a powerful numerical technique since it has the flexibility to model complex geometries with arbitrary shapes and inhomogeneous media. The FEM begins with discretizing the computational domain into smaller elements called finite elements. These finite elements differ for one-, two-, and three-dimensional problems. The next step is to implement the wave equation in a weighted sense over each element, apply boundary conditions and accumulate element matrices to form the overall system of equation [Sadiku, 2009].

5.2 High frequency structure simulator (HFSS™)

Ansoft's High Frequency Structure Simulator (HFSS) is a commercially available and state-of-the-art electromagnetic simulation package. HFSS is one of the industry leading 3D EM software tools for radio frequency (RF) applications. It employs the finite element method (FEM) to simulate any arbitrary three-dimensional structure by solving Maxwell's equations based on the specified boundary conditions, port excitations, materials, and the particular geometry of the structure [HFSS™, v10].

6. The Stepped-Trapezoidal Patch Antenna

6.1 Overview

A novel planar patch antenna with a circular-notch cut fed by a simple microstrip line is proposed and described. It is designed and fabricated for UWB wireless communications and applications over the band 3.1 - 10.6 GHz. This antenna is composed of an isosceles trapezoidal patch with the circular-notch cut and two transition steps as well as a partial ground plane. Because of its structure, we have called it "the stepped-trapezoidal patch antenna" [Alshehri, et al., 2008]. To obtain the UWB bandwidth, we use many bandwidth enhancement techniques: the use of partial ground plane, adjusting the gap between radiating element and ground plane technique, using steps to control the impedance stability and a notch cut technique. The notch cut from the radiator is also used to miniaturize the size of the planar antenna. The measured -10 dB return loss bandwidth for the designed antenna is about 116.3% (8.7 GHz). The proposed antenna provides an acceptable radiation pattern and a relatively flat gain over the entire frequency band. the design details and related results are presented and discussed in the following subsections.

6.2 Antenna design

First, the substrate is chosen to be Rogers RT/Duroid 5880 material with a relative permittivity ε_r=2.2 and a thickness of 1.575 mm. Second, the radiator shape is selected to be trapezoidal since it can exhibit a UWB characteristic. Next, the initial parameters are calculated using the following empirical formula reported in [Evans & Amunann, 1999] after adding the effect of the substrate:

$$f_L(GHz) = \frac{904}{(4\pi h + W + W_1)} \tag{1}$$

Where:
f_L: the frequency corresponding to the lower edge of the bandwidth for the trapezoidal sheet.
W and W_1: the width of the trapezoidal patch bases.
h: the height of the trapezoidal patch.

The dimensions are expressed in mm. This formula is used to predict the lower edge frequency of the bandwidth for the trapezoidal sheet suspended in the space over the ground plane. It is accurate to +/- 9 % for frequencies in the range 500 MHz to 6 GHz. In our design, the sheet will be a patch printed on substrate, so, the effect of the substrate has to be incorprated to the formula. After adding it, the formula becomes:

$$f_L (GHz) = \frac{904}{(4\pi h + W + W_1)\sqrt{\varepsilon_{reff}}} \tag{2}$$

Where the effective relative permittivity ε_{reff} can be calculated using:

$$\varepsilon_{reff} = (\varepsilon_r + 1)/2 \tag{3}$$

Where
ε_r: the relative permittivity of the substrate

Since the antenna is designed for UWB, it has to operate over 3.1 - 10.6 GHz. Therefore, the lower edge frequency at which the initial parameters will be calculated is 3.1 GHz. Initially, the antenna consists of an isosceles trapezoidal patch and partial ground plane etched on opposite sides of the substrate. The radiator is fed through a microstrip line with 50-Ω characteristic impedance. After setting up the configuration of the antenna, determining the initial parameters and fixing the lower frequency, the simulation is started to confirm the calculated parameters. Then, several bandwidth enhancement techniques are applied to widen the bandwidth and to obtain the UWB performance. These techniques are: adjusting the gap between radiating element and ground plane technique, using steps to control the impedance stability and the notch cut technique. It used after studying the current distribution and found out that the current distributions before and after the cut are approximately the same. Also, the notch cut from the radiator is used to miniaturize the size of the planar antenna. Figure 2 illustrates the final geometry of the printed antenna as well as the Cartesian coordinate system.
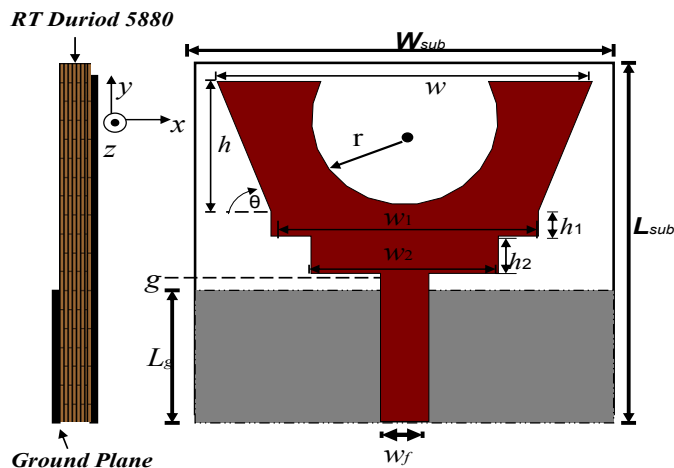
Fig. 2. The geometry of the stepped-trapezoidal patch antenna

It consists of an isosceles trapezoidal patch with notch cut and two transition steps and a partial finite-size ground plane. The Cartesian coordinate system (x,y,z) is oriented such that the bottom surface of the substrate lies in the x-y plane. The antenna and the partial ground plane are etched on opposite sides of the Rogers RT/Duroid 5880 substrate. The substrate

size of the proposed antenna is 30 × 30 mm². The dimensions of isosceles trapezoidal patch are w=28 mm, w_1=20 mm and h=10.5 mm. The first transition step of $w_1 \times h_1$ = 20 mm × 2 mm and second transition step of $w_2 \times h_2$ = 14 mm × 3 mm are attached to the isosceles trapezoidal patch. To reduce the overall size of the printed antenna and to get a better impedance match, the circular-shaped notch with radius r =7 mm is symmetrically cut in the top middle of the isosceles trapezoidal radiator. The shape of the partial ground plane is selected to be rectangular with dimensions of 11 × 30 mm². The radiator is fed through a microstrip line having a length of 12 mm and width w_f =3.6 mm to ensure 50-Ω characteristic impedance with a feed gap of g = 1 mm.

6.3 Parametric study

The parametric study is carried out to optimize the antenna and provide more information about the effects of the essential design parameters. The antenna performance is mainly affected by geometrical and electrical parameters, such as the dimensions related to the notch cut and the two transition steps.

(a) Notch cut

The circular-shaped notch cut is described by its radius and the location of its center. Both parameters are studied. The effect of varying the notch radius on the impedance matching is depicted in Figure 3. When the radius is increased, the entire band is highly affected, especially the middle and higher frequencies experience higher mismatch levels. It is obviously observed that the notch can be used to reduce the size of the radiator provided that the current distribution has low density in the notch part. On the other hand, when the center of the notch moves in the upper side of the patch, the entire band is slightly influenced. In general, the notch cut parameters affect the impedance matching to a certain extent.
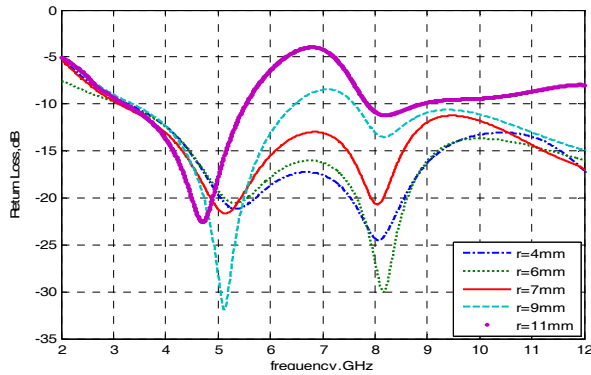
Fig. 3. Effects of notch cut radius

(b) Transition steps

The effects of the two transition steps are studied. They have great impact on the matching impedance for the whole band. For example, the effect of the width of the second step is depicted in Figure 4. From the plot, the step width greatly affects the entire band, especially at the high frequencies range, because the two steps influence the coupling between the

radiator and the ground plane. Thus, by adjusting the steps parameters, the impedance bandwidth can be enhanced. In Figure 6, it is clear that a net improvement on the antenna bandwidth is obtained when the two transitions steps are used.
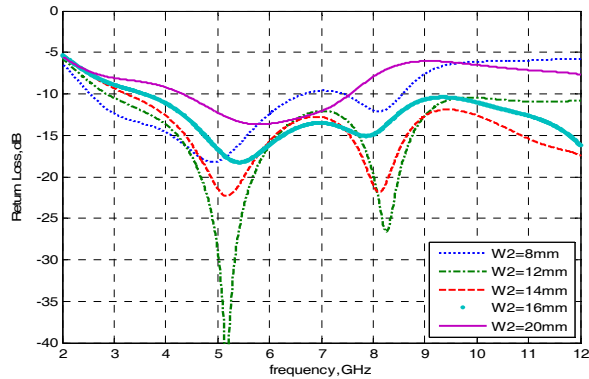
Fig. 4. Effects of step width

6.4 Results and discussion

After taking into account the design considerations described on antenna structure, current distributions and parametric study done to optimize the antenna geometry, the optimized antenna is constructed as shown in Figure 5. Then, the antenna is experimentally tested to confirm the simulation results. The simulated and measured return loss and radiation patterns are presented. Also, the simulated gain is provided.

(a) Front view   (b) Back view
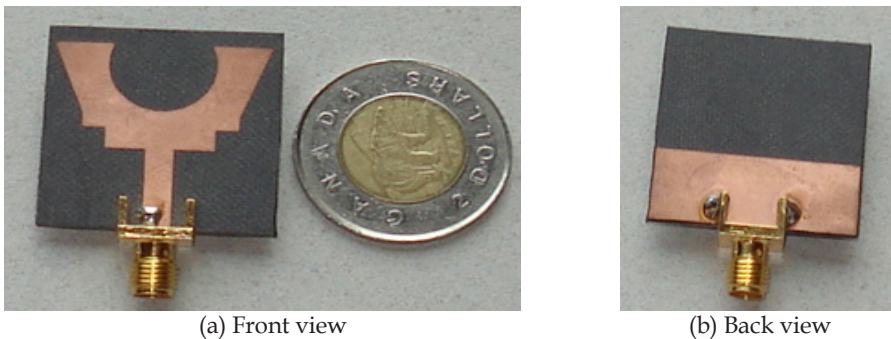
Fig. 5. The prototype of the stepped-trapezoidal patch antenna

(a) Return loss

The return loss (S_{11}) of the proposed antenna is measured. As depicted in Figure 6, the measured and simulated results are shown for comparison and indicate a reasonable agreement. In fact, the simulated return loss of the antenna is found to remain below -10 dB beyond 12 GHz but that range of frequencies is omitted in Figure 6 since it is far out of the allocated bandwidth for UWB communications under consideration. The measured -10 dB return loss bandwidth of the antenna is approximately 8.7 GHz (3.13 - 11.83 GHz). Excellent

performance is obtained since the measured return loss is very close to the simulated one in most range of the frequency band. The measured return loss shows that the antenna is capable of supporting multiple resonance modes, which are closely distributed across the spectrum. Therefore, the overlapping of these resonance modes leads to the UWB characteristic.
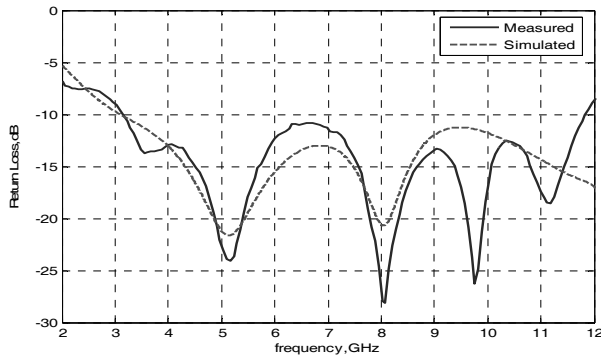
Fig. 6. The simulated & measured return loss

(b) Antenna radiation patterns

The radiation characteristics of the proposed antenna are also investigated. The two dimensional radiation patterns presented here is taken at two sets of principal cuts, $\Theta=0°$ and $\Phi=90°$. Referring to the coordinate system attached to the antenna geometry in Figure 2, the H-plane is the xz-plane and the E-plane is the yz-plane. Figures 7 and 8 illustrate the simulated and measured H-plane and E-plane radiation patterns respectively at 3.5 and 9.5 GHz. In general, the simulated and measured results are fairly consistent with each other at most of the frequencies but some discrepancies are noticed at higher frequencies, especially in the E-plane. These discrepancies are most likely a result of the cable leakage current on the coaxial cable that is used to feed the antenna prototype in the measurements [Kwon & Kim, 2006]. This leakage current is known to be frequency sensitive as well. Also, intrinsic noise within the anechoic chamber may contribute to these discrepancies.

Nevertheless, an analysis of the radiation pattern results shows that the proposed antenna is characterized by omni-directional patterns in the H-plane for all in-band frequencies, as in Figure 7. The measured *H*-plane patterns follow the shapes of the simulated ones well, except at 9.5 GHz where there is little difference.

For the E-plane patterns, Figure 8 shows that they form a figure-of-eight pattern for frequencies up to 7.5 GHz but at 9.5 GHz the shape changes. However, the measured *E*-plane patterns generally follow the simulated ones well. In general, the stepped-trapezoidal patch antenna shows an acceptable radiation pattern variation in its entire operational bandwidth since the degradation happens only for a small part of the entire bandwidth and it is not too severe.

(a) H-plane at 3.5 GHz   (b) H-plane at 9.5 GHz

Fig. 7. The simulated and measured radiation patterns in the H-plane

(a) E-plane at 3.5 GHz   (b) E-plane at 9.5 GHz
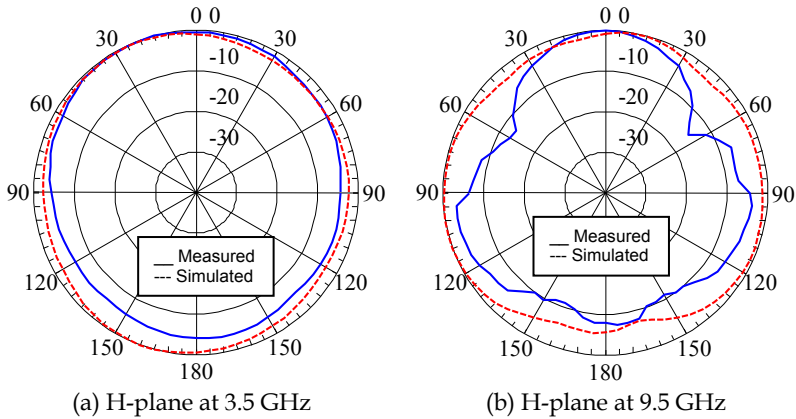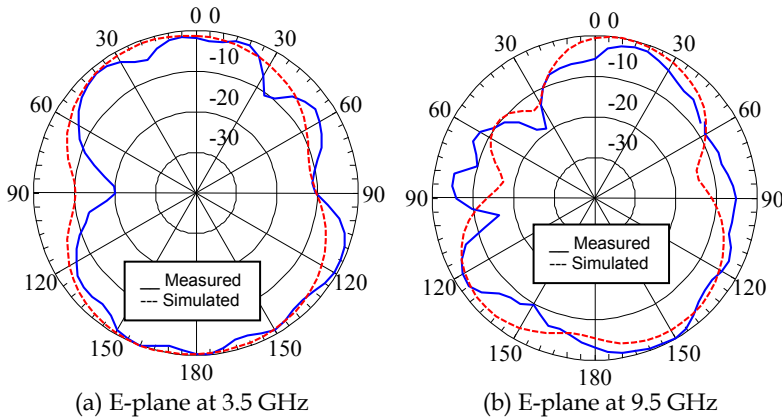
Fig. 8. The Simulated and measured radiation patterns in the E-plane

(c) Antenna gain

The gain of the proposed antenna is also found to be suitable for the UWB communications and applications. It is greater than 2.9 dBi for all in-band frequencies and varies from 2.9 dBi to 5.2 dBi over the operating frequency range, resulting in the maximum gain variation of 2.3 dB.

7. The Double-Beveled Patch Antenna

7.1 Overview

A novel planar patch antenna with a notch-cut fed by a simple microstrip line is proposed and described. It is designed and fabricated for UWB wireless communications and applications under the band (3.1-10.6 GHz). This antenna is composed of a symmetrical double-beveled planar patch antenna with notch cut fed by a microstrip line folded on a partial ground plane. Because of its structure, we have called it "the Double-Beveled Patch

Antenna" [Alshehri & Sebak, 2008]. To obtain UWB bandwidth, we use several bandwidth enhancement techniques: the use of partial ground plane, adjusting the gap between radiating element and a ground plane technique, the use of bevels technique and a notch cut technique used also to reduce the size of the planar antenna. A parametric study is numerically carried out on the important geometrical parameters to understand their effects on the proposed antenna and therefore optimize its performance. The measured -10 dB return loss (VSWR<2) bandwidth is about 123.8% (9.74 GHz). The proposed antenna provides an acceptable radiation pattern and a relatively flat gain over the entire frequency band. The measured and simulated results for both bandwidth and radiation pattern show a very reasonable agreement. In the following subsections, the design details and the related results are presented and discussed.

7.2 Antenna design

First, the substrate is chosen to be Rogers RT/Duroid 5880 material with a relative permittivity ε_r = 2.2 and a thickness of 1.575 mm. Second, the radiator shape is selected to be rectangular. Next, the initial parameters are calculated using the empirical formula reported in [Agrawall, et al., 1998] after adding the effect of the substrate:

It is found that the frequency corresponding to the lower edge of the bandwidth of the monopole antenna can be predicted approximately by equating the area of the planar configuration to that of a cylindrical wire and given by:

$$2\pi r l = hW \tag{4}$$

So, the resonant frequency is given by:

$$f_L(GHz) = \frac{c}{\lambda} = \frac{30 \times 0.24}{l + r} \tag{5}$$

Where:
f_L: the frequency corresponding to the lower edge of the bandwidth.
C: the light speed.
λ : the wavelength
l: the height of the cylindrical wire which is same as that of planar configuration height
r: the equivalent radius of the cylindrical wire
W: the width of the rectangular patch.
h: the height of the rectangular patch.

The dimensions are expressed in centimeters. This simple formula is used to predict the lower edge frequency of the bandwidth for the monopole suspended in the space over the ground plane. It is accurate to +/- 8 %. In our design, the sheet will be a patch printed on the substrate, so, the effect of the substrate has to be included to the formula. After considering it, the formula becomes:

$$f_L(GHz) = \frac{30 \times 0.24}{(l + r)\sqrt{\varepsilon_{reff}}} \tag{6}$$

where the effective relative permittivity ε_{reff} can be calculated using Equation 3.

Since the antenna is designed for UWB, it has to operate over 3.1 - 10.6 GHz. Therefore, the lower edge frequency at which the initial parameters will be calculated is 3.1 GHz. Initially, the antenna consists of a rectangular patch and partial ground plane etched on opposite sides of the substrate. The radiator is fed through a microstrip line with 50-Ω characteristic impedance. After setting up the configuration of the antenna, determining the initial parameters and fixing the lower frequency, the simulation is performed to confirm the calculated parameters. Then, several bandwidth-enhancement techniques are applied to widen the bandwidth and obtain UWB performance. These techniques are: adjusting the gap between radiating element and ground plane technique, the bevels technique and notch cut technique used after studying the current distribution as will be discussed later.

Figure 9 illustrates the geometry of the printed antenna as well as the Cartesian coordinate system. It consists of a symmetrical double-beveled patch with notch cut and a partial ground plane. The Cartesian coordinate system (x,y,z) is oriented such that the bottom surface of the substrate lies in the x-y plane. The antenna and the partial ground plane are oppositely etched on the Rogers RT/Duroid 5880 substrate. The substrate size of the proposed antenna is 40 × 31 mm².
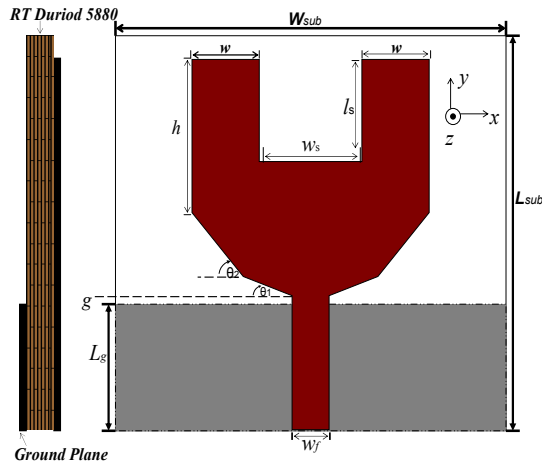
Fig. 9. The geometry of the double-beveled patch antenna

The parameters of the symmetrical double-beveled patch are w=6.5 mm, h=12 mm, θ_1 =17.5° (the angle of the first bevel) and θ_2 =45° (the angle of the second bevel). To reduce the overall size of the printed antenna and to get better impedance matching, a rectangular-shaped notch with dimensions of $l_s \times w_s$ = 8 mm × 10 mm is symmetrically cut in the top middle of the radiator. The shape of the partial ground plane is rectangular with dimensions of 10 × 40 mm². The radiator is fed through a microstrip line having a length of 10.5 mm and width w_f =3.6 mm to ensure 50-Ω input impedance with a feed gap of g = 0.5 mm. The 50 Ω-microstrip line is printed on the same side of the substrate as the radiator.

7.3 Current distribution

The current distribution is studied. The simulated current distributions of the initial antenna geometry before cutting the region of low current density at 3.5 and 9.5 GHz (as examples) are shown in Figure 10 (a) and (b) respectively. The current is mainly concentrated on the bottom portion of the patch with very low density toward and above the center and it is distributed along the edges of the patch, except the top edge, for all frequencies. Thus, it can conclude that the region of low current density on the patch is not that important in the antenna performance and could therefore be cut out. Consequently, a rectangular section with dimensions of $l_s \times w_s$ = 8 mm × 10 mm is symmetrically cut out from the top middle of the rectangular radiator to eliminate a region of low current density as shown in Figure 9. After this cut, the current distributions at 3.5 GHz and 9.5 GHz (as examples) are depicted in Figure 10 (c) and (d), respectively. It is observed that the current distributions in this case are approximately the same as before the cut. As a result of this cut, the size of the antenna is reduced and has lighter weight, which is very desirable for more degree of freedom in design and possibly less conductor losses.

(a) at 3.5 GHz

(c) at 3.5 GHz

(b) at 9.5 GHz

(d) at 9.5 GHz
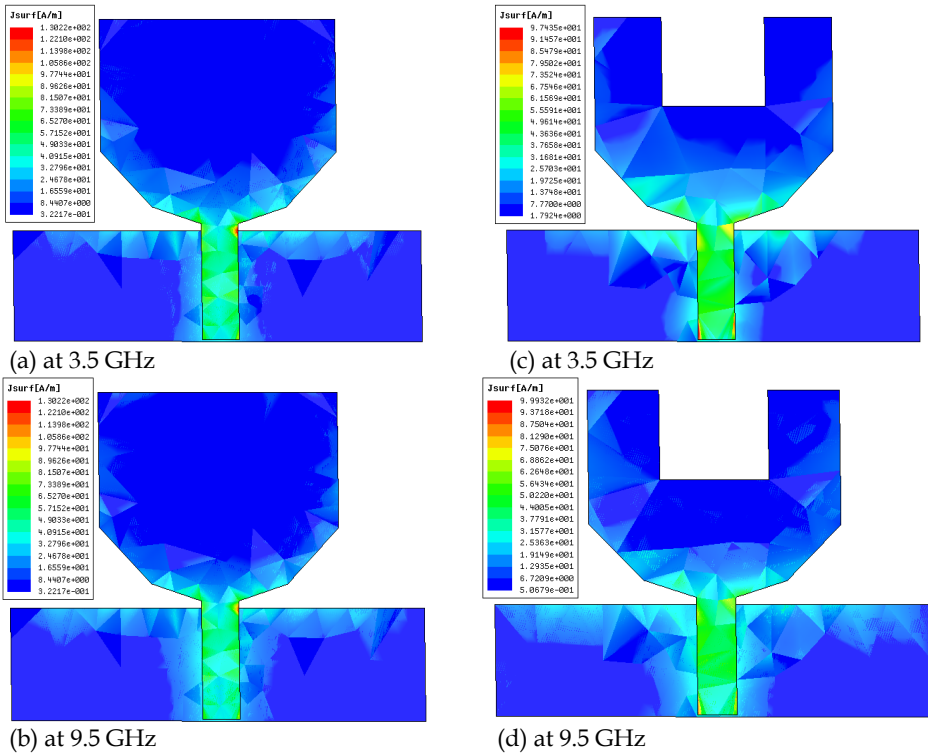
Fig. 10. The current distributions

7.4 Parametric study

The parametric study is done to optimize the antenna. Its performance is mainly affected by geometrical parameters, such as the dimensions related to the notch cut and the bevels.

(a) Notch cut

The effect of the rectangular-shaped notch dimensions (l_s, w_s) on the return loss is studied. It is observed that the width of the notch has a major effect on the impedance matching over the entire frequency range, as shown in Figure 11. The lower edge frequency of the bandwidth is shifted to higher frequencies once the width increases. Also, the middle and higher frequencies are affected with higher mismatch levels. On the other hand, the length of the notch slightly influences the lower edge frequency. It is also observed that the notch can be used to reduce the size of the radiator, as explained earlier using the current distribution.
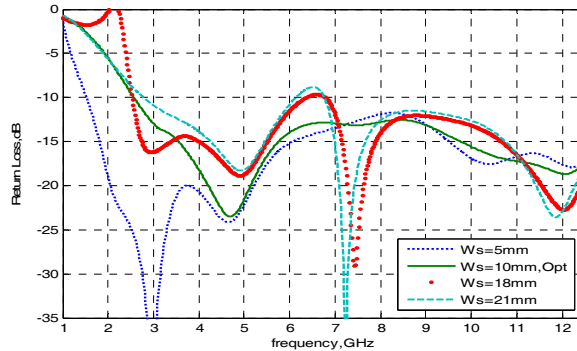
Fig. 11. Effects of the width of notch cut

(b) Bevels

The double bevels dimensions influence the matching impedance for the whole band, especially at high frequencies. The high frequencies can be controlled and the entire band can be enhanced by adjusting the bevel angles. By varying the angle of the first bevel (θ_1), the low and middle frequencies are highly influenced. As shown in Figure 12, by varying the angle of the second bevel (θ_2), the whole band is affected especially at middle and high frequencies. Thus, using two progressive bevels provides more degree of freedom and by adjusting them, the bandwidth will be widened as well as excellent level of matching can be achieved.
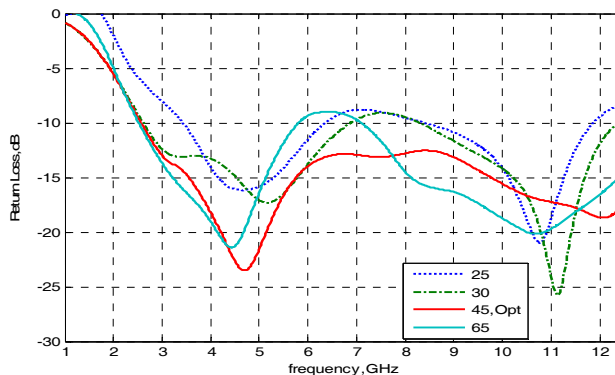
Fig. 12. Effects of second bevel angle

7.5 Results and discussion

After taking into account the design considerations described on antenna structure, current distributions and parametric study done to optimize the antenna geometry, the optimized antenna is constructed as shown in Figure 13 using the optimum values as mentioned earlier . Then, the antenna is experimentally tested to confirm the simulation results. The simulated and measured VSWR is presented as well as the simulated and measured radiation patterns in principle planes. Also, the simulated gain is provided.

(a) Front view   (b) Back view
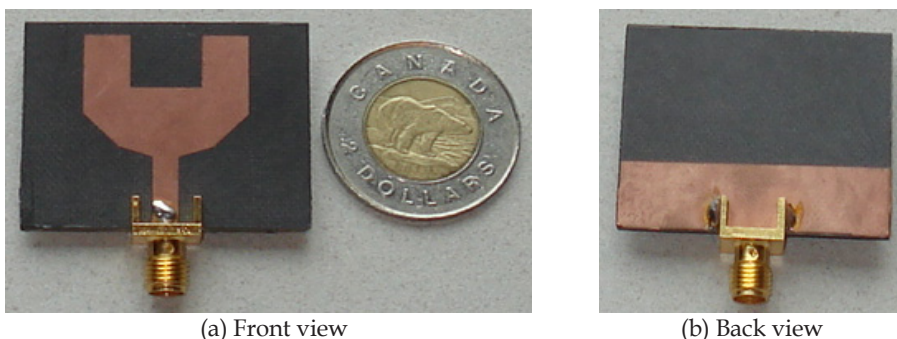
Fig. 13. The prototype of the double-beveled patch antenna

(a) VSWR

The VSWR of the proposed antenna is measured as depicted in Figure 14. The measured -10 dB return loss (VSWR<2) bandwidth of the antenna is approximately 9.74 GHz (3.00-12.74 GHz) and the antenna shows stable behaviors over the band. Thus, the measurement confirms the UWB characteristic of the double-beveled patch antenna as predicted in the simulation.
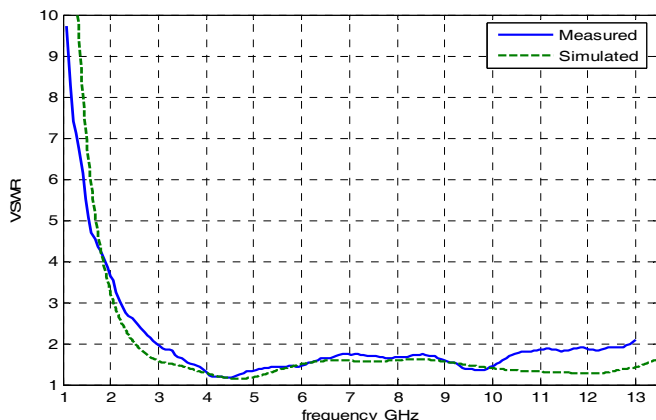
Fig. 14. Simulated & measured VSWR

(b) Antenna radiation patterns

The radiation characteristics of the proposed antenna are also investigated. Figures 15 and 16 illustrate the simulated and measured H-plane and E-plane radiation patterns

respectively at 3.5, 5.5, 7.5 and 9.5 GHz. In genral, the simulated and measured results are fairly consistent at most of the frequencies but some discrepancies are noticed at higher frequencies especially in the E-plane. Nevertheless, the proposed antenna is characterized by omni-directional patterns in the H-plane for all in-band frequencies as in Figure 15. For the E-plane patterns, Figure 16 shows that the simulated ones at low frequencies form figure-of-eight patterns but at high frequencies, there are dips, especially at 9.5 GHz. In general, the double-beveled patch antenna shows an acceptable radiation pattern variation in its whole operational bandwidth since the degradation happens only for a small part of the entire bandwidth and it is not too drastic.

(a) H-plane at 3.5GHz

(b) H-plane at 5.5GHz

(c) H-plane at 7.5GHz

(d) H-plane at 9.5GHz

Fig. 15. The simulated and measured radiation patterns in the H-plane

(a) E-plane at 3.5GHz   (b) E-plane at 9.5GHz
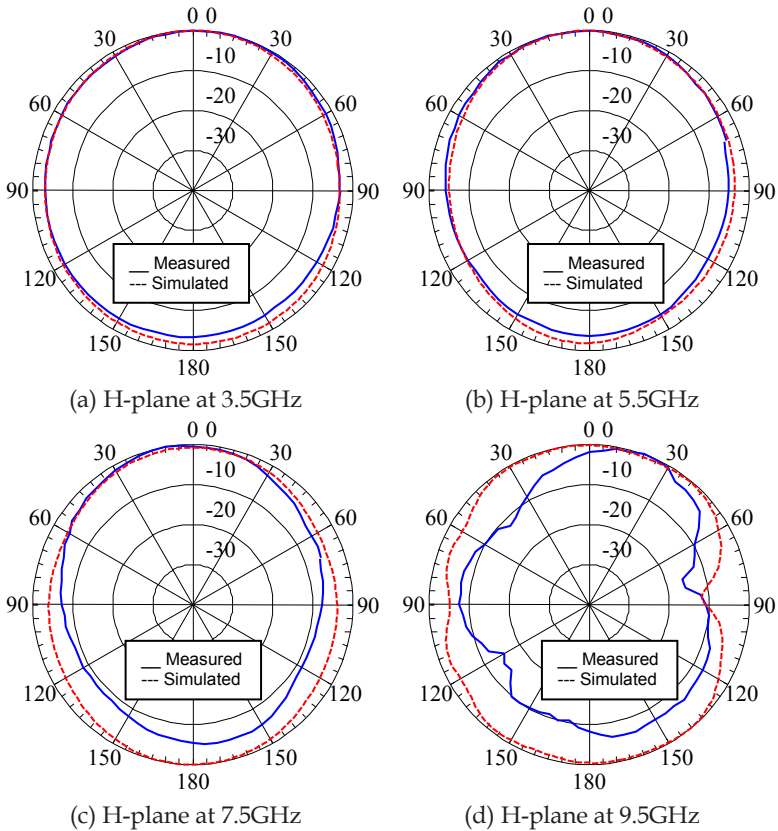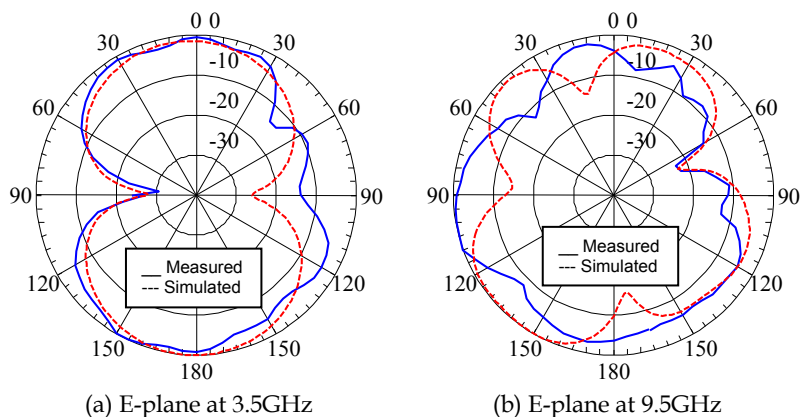
Fig. 16. The simulated and measured radiation ratterns in the E-plane

(c) Antenna gain

The gain versus frequency of the proposed antenna is also found to be suitable for the UWB communications and applications. The simulated antenna gain versus frequency is shown in Figure 17. It is greater than 3.4 dBi for all in-band frequencies and varies from 3.4 dBi to 6.1 dBi over the operating frequency range, resulting in the maximum gain variation of 2.7 dB.
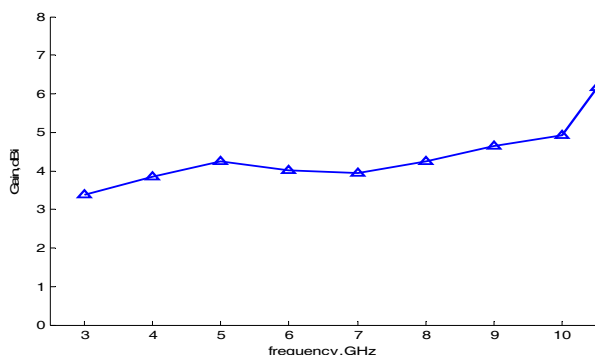
Fig. 17. Simulated gain

8. Conclusion

An overview on Ultra Wideband wireless communications is given. Two novel, small, low-profile, microstrip-fed printed UWB antennas are analyzed, designed and implemented to satisfy UWB technology requirements. The focus is on UWB frequency domain characteristics such as the far field radiation patterns, bandwidth and gain. The antenas provide excellent performance in the entire operational bandwidth. Because of their low cost, light weight and ease of implementation, these printed designs are desired in UWB wireless communication systems and applications, especially in portable devices and indoor applications such as WPAN. These antennas are namely: the stepped-trapezoidal patch antenna and the double-

beveled patch antenna. Both provides a nearly omni-directional radiation pattern and a relatively flat gain over the entire frequency band with a maximum variation of 2.3 dB for first one and 2.7 dB for the second one. Both antennas offer reduced patch size, more degree of freedom for design, extra space that could accommodate other RF circuit elements. However, the effect Analysis of the notch cut show that within a certain limit of the cutout size, the radiation properties do not change drastically. But beyond that limit, the notch cut highly affects radiation patterns in the entire operational bandwidth.

UWB systems occupy huge operational bandwidth and often utilize very short pulses for data transmission. Therefore, an appropriate time domain performance is a key requirement for UWB antennas. Accordingly, investigations and analysis will be carried out on the effect of the proposed antennas on the transmitted pulse to hence improve the time domain behavior by optimizing the antenna designs.

9. References

Agrawall, N.; Kumar, G. & Ray, K. (1998). Wide-Band Planar Monopole Antennas, *IEEE Transactions on Antennas and Propagation*, vol. 46, No., February 1998, (294-295), 0018-926X

Alshehri A. (2008). *Novel Ultra Wideband Antennas for Wireless Systems*, M. A. Sc. Thesis, Concordia University, Canada

Alshehri, A. & Sebak, A. (2008). A Novel UWB Planar Patch Antenna for Wireless Communications, *IEEE International Symposium on Antennas and Propagation*, pp. 1-4, 978-1-4244-2041-4, USA, July 2008, IEEE, San Diego

Alshehri,A.; Sebak, A. & Denidni, T. (2008). A Novel UWB Stepped-Trapezoidal Patch Antenna for Wireless Communications, *The IASTED International Conference on Antennas, Radar and Wave Propagation (ARP 2008)*,pp. 27-31, 978-0-88986-735-2, USA, April 2008, ACTA Press, Calgary

Ammann, M. & Chen, Z. (2003). A Wide-Band Shorted Planar Monopole with Bevel, *IEEE Transactions on Antennas and Propagation*, Vol. 51, No.4, April 2003, (901-903), 0018-926X

Ammann, M. & Chen, Z. (2004). An Asymmetrical Feed Arrangement for Improved Impedance Bandwidth of Planar Monopole Antennas, *Microwave and Optical Technology Letters*, Vol. 40, No. 2, January 2004, (156-158)

Ammann, M. & Chen, Z. (2003). Wideband Monopole Antennas for Multi-Band Wireless Systems, *IEEE Antennas and Propagation Magazine*, vol. 45, No. 2, April 2003, (146-150), 1045-9243

Ammann, M. (2001), Control of the Impedance Bandwidth of Wideband Planar Monopole Antennas Using A Beveling Technique, *Microwave and Optical Technology Letters*, Vol. 30, No. 2July 2001, (229-232)

Arslan, H.; Chen, Z. & Benedetto, M. (2006). *Ultra Wideband Wireless Communication,* John Wiley & Sons, Inc., 0471715212, New Jersey

Azenui, N. (2007). *Miniaturized Printed Circuit Antennas for Multi- and Ultra-Wide Band Applications*, PHD Thesis, University of Illinois, USA

Balanis, C. (2005). *Antenna Theory Analysis and Design,* John Wiley & Sons, Inc., 047166782X, New Jersey

Barrett, T. (2000). History of Ultra Wideband (UWB) Radar & Communications: Pioneers and Innovators," *Progress in Electromagnetics Symposium 2000*, USA, July 2000.

Bataller, M.; Fabrés, M.; Daviu E. et al. (2006). Overview of Planar Monopole Antennas for UWB Applications, *Proceeding of EUCAP 2006*, ESA SP-626, November 2006, France

Bhartia, P.; Bahl, I.; Garg, R.; et al. (2000). *Microstrip Antenna Design Handbook*, Artech House, 0890065136, USA

Cai, A.; See, T. & Chen, Z. (2005). Study of Human Head Effects on UWB Antenna, *IEEE International Workshop on Antenna Technology (iWAT)*, 0-7803-8842-9, March 2005, pp. 310–313, IEEE

Chen, D. & Chen, H. (2004). A CPW-Fed Dual-Frequency Monopole Antenna, *IEEE Transactions on Antennas and Propagation*, vol. 52, No. 4, April 2004, (978-982), 0018-926X

Chen, Z.; Ammann, M. & Chia, M. (2003). Broadband Square Annular Planar Monopoles," *Microwave and Optical Technology Letters*, vol. 36, No. 6, March 2003, (449-454)

Chen, Z.; Ammann, M.; Qing X., et al. (2006). Planar Antennas, *IEEE Microwave Magazine*, Vol. 7, No. 6, December 2006, (63-73), 1527-3342

Chen, Z.; See, T. & Qing, X. (2007). Small Printed Ultra Wideband Antenna with Reduced Ground Plane Effect, *IEEE Transactions on Antennas and Propagation*, vol. 55, No. 3 February 2007, (383-388), 0018-926X

Choi, S.; Lee, H. & Kwak, K. (2006). Clover-Shaped Antenna for Ultra-Wideband Communications, *Microwave and Optical Technology Letters*, Vol. 48, No. 10, July 2006 (2111-2113).

Choi, S.; Park, J.; Kim, S. et al. (2004). A New Ultra-Wideband Antenna for UWB Applications, *Microwave and Optical Technology Letters*, Vol. 40. No. 5, March 2004, (399-401)

Daviu, E.; Fabres, M.; Bataller, M. et al. (2003). Wideband Double-Fed Planar Monopole Antennas, *Electronics Letters*, Vol. 39, November 2003, (1635-1636), 0013-5194

Dotto, K. (2005) *Development of a Novel Ultra-Wideband Antenna and Prototype Scanner for Detection and Location of Voids in Wood*, PHD Thesis, The University of British Columbia

Dyson, J. (1959). The Unidirectional Equiangular Spiral Antenna, *IEEE Transactions on Antennas Propagation*, Vol. 7, No. 4, October 1959, (329-334), 0096-1973

Evans, J.; Amunann, M. (1999), Planar Trapezoidal and Pentagonal Monopoles with Impedance Bandwidths in Excess of 10:1, *IEEE International Antenna and Propagation Symposium*, pp.1558-1561, 0-7803-5639-x, Orlando, FL, July 1999, IEEE, Orlando

Federal Communication Commission (2002). *First Report and Order, Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission System*, FCC 02 48

HFSS™, v10, Ansoft Corporation Software, Pittsburgh, PA, USA

Huang, C. & Hsia, W. (2005). Planar Elliptical Antenna for Ultrawideband Communications, *Electronics Letters*, Vol. 41, No. 6, March 2005, (296-297), 0013-5194

Kwon, D. & Kim, Y. (2006). Suppression of Cable Leakage current for Edge-Fed Printed Dipole UWB Antennas Using Leakage-Blocking Slots, *IEEE Antennas and Wireless Propagation Letters*, vol. 5, No. 1 ,December 2006, (183-186), 1536-1225

Lee, S.; Park, J. & Lee J. (2005). A Novel CPW-Fed Ultra-Wideband Antenna Design, Microwave Optical Technology Letters, Vol. 44, No. 5, January 2005, (393-396)

Lee, E.; Hall, P. & Gardner, P. (1999). Compact Wideband Planar Monopole Antenna, *Electronics Letters*, Vol. 35, December 1999, (2157-2158), 0013-5194

Lee, F.; Wentzloff,D. & Chandrakasan, A. (2004). An Ultra-Wideband Baseband Front-End, *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp.493-496, 0-7803-8333-8, USA, June 2004, IEEE, Cambridge

Liang, J. (2006). *Antenna Study and Design for Ultra Wideband Communication Applications*, PHD Thesis, University of London, United Kingdom

Lu, G. (2006). A*ntenna and Synchronization Design Issues for the Ultra-Wideband System*s, PHD Thesis, the State University of New Jersey, USA,

Muscat, A. & Parini, C. (2001). Novel Compact Handset Antenn, IEEE *11th International Conference on Antennas and Propagation*, pp. 336-339, 0-85296-733-0, Manchester, UK, April 2001, IEEE

Oppermann, I.; Hamalainen, M. & Iinatti, J. (2004). *UWB Theory and Applications,* John Wiley & Sons, Ltd, 0470869178, New Jersey

Powell, J. (2004). *Antenna Design for Ultra Wideband Radio*, M. A. Sc. Thesis, Massachusetts Institute of Technology, USA

Ross, G. (1968). A Time Domain Criterion for the Design of Wideband Radiating Elements, *IEEE Transactions on Antennas Propagation*, vol. 16, No. 3, May 1968, (355- 356), 0018-926X

Ross, G. (1973). Transmission and Reception System for Generating and Receiving Base-Band Duration Pulse Signals for Short Base-Band Pulse Communication System, April 1973, U.S. Patent no. 3728632

Rudge, A.; Milne, K.; Olver, A.; (1982). *The Handbook of Antenna Design,* Peter Peregrinus Ltd., 0906048826, UK

Rumsey, V. (1957). Frequency Independent Antennas, *IRE National Convention Record*, Vol. 5, March 1957, (114-118)

Sadiku, M. (2009). *Numerical Techniques in Electromagnetics,* CRC Press LLC., 142006309X USA

Schantz, H. (2003). A Brief History of Ultra-Wideband Antennas, *IEEE Conference on UWBST*, pp. 209-213, 0-7803-8187-4, November 2003

Siriwongpairat, W. (2005). *Cross-Layer Design for Multi-Antenna Ultra-Wideband Systems,* PHD Thesis, University of Maryland, USA

Siwiak, K. & McKeown, D. (2004) *Ultra-Wideband Radio Technology,* John Wiley & Sons, Ltd., 0470859318, New Jersey

Tsai, M. & Wang, H. (2004). A 0.3-25 GHz Ultra-Wideband Mixer Using Commercial 0.18-/μm CMOS Technology, *IEEE Microwave and Wireless Components Letters*, Vol. 14, No. 14 November 2004, (522-524), 1531-1309

Wentzloff, D.; Blazquez R. ; Lee F. , et al. (2005), System Design Considerations for Ultra-Wideband Communication, *IEEE Communications magazine*, Vol. 43, August 2005, (114-121), 0163-6804

Wong, K.; Chi, Y.; Su C.; et al. (2005). Band-Notched Ultra-Wideband Circular-Disc Monopole Antenna with an Arc-Shaped Slot, *Microwave and Optical Technology Letters*, vol. 45, No. 3, May 2005, (188-191)

Yang, L.; Giannakis, G. (2004). Ultra-Wideband Communications: An Idea Whose Time Has Come, IEEE Signal Processing Magazine, Vol. 21, No. 6, November 2004, (26-54), 1053-5888

Micromachined high gain wideband antennas for wireless communications

Sumanth K. Pavuluri, Changhai Wang and Alan J. Sangster
Heriot Watt University
Edinburgh, EH14 4AS, UK

1. Introduction

The seemingly insatiable and growing demand for compact, multi-function, multi-frequency electronic systems for communications and other applications, is continuing to drive the search for devices offering more and more bandwidth. There is growing need for broadband high gain communication systems in the X band range of frequencies (8 - 12 GHz) for terrestrial broadband communications and networking as well as for radar applications. Similarly, direct broadcast satellite (DBS) and various other applications in the K_u band (10 - 14 GHz) such as radio astronomy service, space research service, mobile service, mobile satellite service, radio location service (radar), amateur radio service, and radio navigation may require embedded antenna systems at different bands. It would be ideal if efficient, broadband and cost effective planar microstrip based antenna and antenna array devices could be designed to provide coverage of all these bands. In addition systems aimed at UWB (Ultra Wide Band) operation need efficient very wideband antenna devices.

For these high frequency systems, compact size and high performance can usually be achieved by fabricating the antenna onto a low dielectric constant material and integrating it with the remaining circuitry implemented on a high dielectric constant substrate in neighbouring regions in the same package. This trend has serious implications for antennas, where these are required to be embedded within the system package, such as a mobile phone. Systems operating in the microwave and millimetre-wave frequency bands offer the possibility of high levels of integration of individual devices in high density layouts. The most compact circuit designs are invariably achieved by employing high dielectric constant substrates, but this is a requirement which is essentially incompatible with the needs of an embedded planar antenna. Such antennas radiate most efficiently when fabricated onto substrates which exhibit low dielectric constant (Papapolymerou et al., 1998). While it is not impossible to fabricate microstrip or coplanar circuits, together with planar antennas on the same high permittivity silicon substrate, antenna gain and efficiency will inevitably be very poor.

Various schemes have been suggested, in recent years, aimed at overcoming the opposing substrate requirements of circuits and antennas. These largely involve the use of layered materials with high and low permittivities in adjacent layers (Chen, 2008). However such

methods tend to be of quite limited versatility and the trend now is toward selective removal of substrate in the vicinity of the antenna. This can be done by, for example, bulk micromachining an air gap between the planar antenna (usually a conducting patch) and the ground plane (Koul, 2007). The advantages of doing so are as follows:

- Lower effective dielectric constant, hence wider circuit dimensions
- Ease of fabrication and relaxed dimensional tolerances
- Lower attenuation
- Enhanced radiation efficiency in case of antennas
- Eliminating surface waves

Micromachining technology continues to develop, and it is being applied in new ways to embedded antennas to improve their performance. The use of selective lateral etching based on micromachining techniques to enhance the performance of rectangular microstrip patch antennas printed on high-index wafers such as silicon, GaAs, and InP have been developed in the past decade. A novel polymer micromachining based method for achieving high performance, cost effective antennas is described in this chapter.

2. Micromachined antennas

Over the last decade several micromachining techniques have been developed for producing microwave wave and millimeter wave antennas. Devices using these procedures have achieved high performances compared to the conventional patches printed on to relatively high dielectric constant substrates. Various micromachining methods that have been implemented recently are listed in the following sections.

2.1 Silicon micromachining

Silicon micromachining has been employed to fabricate a patch antenna wherein, the silicon material was removed laterally underneath it thus producing a cavity that consists partly of air and partly of substrate (Papapolymerou et al., 1998, Hou et al., 2008, Ojefors et al., 2006, Kratz and Stenmark, 2005). Examples with both equal and unequal thicknesses of air and substrate have been implemented. The micromachined antenna configuration consisted of a rectangular patch centred over the cavity, sized according to the effective index of the cavity region, and fed by a microstrip line. To produce the mixed substrate cavity region, silicon micromachining was used to laterally remove the material from underneath the patch resulting in two separate dielectric regions of air and silicon. The amount of silicon removed varied from 50 to 80% of the original substrate thickness underneath the patch. A cavity model was used to estimate the effective refractive index value below the patch. The walls of the hollowed cavity tend to be, slanted owing to the anisotropic nature of the chemical etching, and this has to be allowed for in the modelling. This antenna has been shown to exhibit superior performance over conventional designs with the bandwidth and the efficiency having been increased by as much as 64% and 28%, respectively.

2.2 Polymer micromachining

Thick photoresist patterning processes can be used to fabricate an air suspended patch antenna either with supporting metallic posts or polymer posts. Antenna structures at

different frequency bands require different air cavity thickness to achieve optimum antenna performance and better impedance matching. Photoresist based polymers such as SU8 and THB151N can be used to obtain ultra thick supporting posts and can also be used as moulds for electroplating metal posts. Various polymer micromachining methods have been implemented in the past (Ryo-ji and Kuroki, 2007). A CPW fed post supported patch antenna has been fabricated on a Corning 7740 glass substrate which had a thickness of 800 µm and a dielectric constant of 4.6. Copper was used for metallization. The feed line of the antenna was patterned with the thick photoresist of AZ9260 and a two-step coating process was performed to form the posts of the antenna with a thick photoresist of THB151N. A simulated antenna gain in the range of 5.6 dBi to 9.0 dBi and the radiation efficiency varying from 92.8 % to 97.4 % were demonstrated for single patch antennas. In the case of a 2 × 1 array patch antenna, the simulated antenna gain and the radiation efficiency were from 5.8 dBi to 11.2 dBi and from 93.6 % to 95.3 %, respectively.

SU8, a widely used negative tone photoresist, has been used to fabricate an elevated patch antenna with micromachined posts of around 800 µm of height. (Pan et al., 2006; Bo et al., 2005) have successfully demonstrated an air-lifted patch antenna fabricated using surface micromachining technology. Both metal posts and polymer posts were used to provide mechanical support, as well as electrical excitation. A -10 dB bandwidth of 7%, centred at 25 GHz, was obtained. The proposed structure is superior to the conventional patch in terms of bandwidth, efficiency and lower side lobe level. While the traditional patch antenna directly printed on substrate usually gives a 3%-5% bandwidth and 70%-80% radiation efficiency, the proposed elevated patch will double the fractional bandwidth and gives a theoretical 97% radiation efficiency. This is achieved by eliminating the substrate loss. Low permittivity spin-on dielectric substrates are efficient for guiding microwaves and millimetre waves (Wang et al., 2005) and they have been used for microwave filters to improve the insertion loss of devices fabricated on silicon substrates (Leung et al., 2002).

2.3 Millimeter wave antennas using low permittivity dielectric substrates and micromachining

Antennas using low permittivity dielectric substrate have wider impedance bandwidth and higher gain when compared with those using ceramic dielectric substrates. Tong et al have presented the simulation and measurement of millimeter-wave CPAs (Coplanar patch antennas) using spin-on low-k dielectric substrate (Tong et al., 1995). The antenna composes of a gold ground plane at the bottom, two layers of BCB dielectric substrate (ε_r = 2.7 and $tan\delta$ = 0.002 @ 20GHz) in the middle and a CPA pattern on the top. The total thickness of the BCB layer is 30 µm. Fluid state BCB is spun onto a 3-inch ground plane coated silicon wafer. The deposition technique is similar to the commonly used photoresist coating technique and the metal CPA pattern is evaporated onto the BCB dielectric layer. The thicknesses of the ground plane and the CPA pattern are both about 1.5 µm. The simulated and measured impedance bandwidths are about 1.2% and 2.6% respectively. The measured resonant frequency of the antenna is 38.3 GHz. Micromachining techniques employing closely spaced holes have been used underneath a microstrip antenna on a high dielectric-constant substrate to synthesize a localized low dielectric-constant environment (ε_r = 2.3) (Gauthier et al., 1997). The holes are drilled using a numerically controlled machine (NCM) and extend at least 3.5 mm from the edge of the antenna in all directions and occupy the full substrate

height. The measured radiation efficiency of a microstrip antenna on a micromachined substrate (Duroid 6010, ε_r = 10.8) increased from 48.3% to 73.3% at 12.8 – 13.0 GHz (including the loss with a 3.3 cm long feed line).

2.4 Integrated chip-size antennas using laser micromachining

Integrated laser micromachined antennas on glass substrates are compact due to reduced size and owing to the relatively high dielectric constant. Mendes et al have reported an on-chip integrated, folded short-patch antenna (Mendes et al., 2006). It consists of three horizontal metal sheets that are electrically connected by two vertical metal walls. All this is embedded in a glass substrate having defined electrical permittivity and dielectric losses. The antenna was designed to operate at 5.1 GHz, a frequency chosen to be inside the 5 – 6 GHz ISM band. The fabricated antenna has dimensions of 4mm×4mm×1mm, a measured operating frequency of 5.05 GHz with a bandwidth of 200 MHz at the return loss of −10 dB and a simulated radiation efficiency of 60%. A method of applying laser micromachining technologies to fabricate a compact, high performance and low-cost 3D monopole antenna was proposed by Huang et al (Huang et al., 2005). The coplanar waveguide (CPW) fed configuration was used owing to its simple structure, wide bandwidth, and the ability of multi-band operation.

2.5 LTCC micromachining

LTCC multilayer technology can be used to build up antenna arrays because it provides the necessary degree of vertical integration for the high-density microwave circuit and a packaging solution such as SIP (System in Package) (Wolff, 2007, Baras and Jacob, 2008, Wi et al., 2007). To optimise the material properties by reducing the relative permittivity of LTCC, a material modulation procedure based on punching air holes into the substrate is performed. Thereby, the relative permittivity of the material is replaced by the effective relative permittivity ε_r of the modulated material (Schuler et al., 2003).

3. Antenna design and modelling

A compact planar antenna which is eminently suitable for embedding within integrated electronic systems is the patch antenna shown in Fig. 1. This type of antenna can be excited in one of four ways (Pozar, 1992, James and Hall, 1989, Bahl and Bhartia, 1980): (a) directly from a microstrip line, (b) electromagnetically from a buried line, (c) directly from a coaxial line probe, and (d) electromagnetically by means of a slot. Of these only (d) permits the implementation of multi-layered formats and consequently the section is directed toward assessing this geometry in relation to the micromachining of such structures.
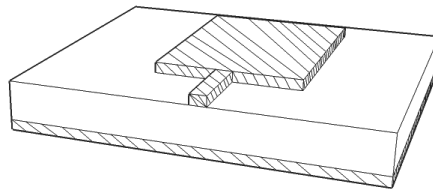
(a) Microstrip feed
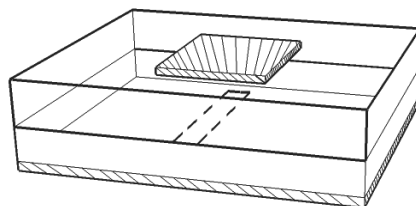
(b) Buried line feed
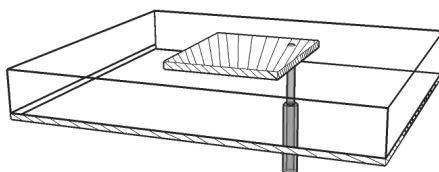
(c) Coaxial feed
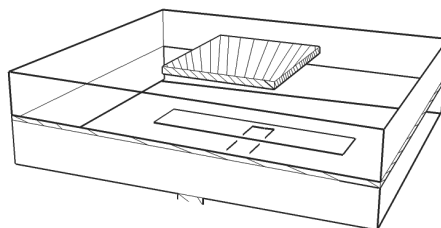
(d) Aperture coupled feed

Fig. 1. Illustration of feeding methods for microstrip antennas.

High gain and wideband micromachined microstrip and aperture fed stacked patch antenna devices have been designed and are described in the following sections. The antenna devices have been modelled using an electromagnetic simulation package. The structures are fabricated in layers and then assembled to obtain 3D devices. The detailed fabrication and assembly processes will be described. These aperture coupled devices are impedance matched for wideband operation. RF measurements show high gain and wideband operation for the devices and the results are in good agreement with that of simulation. The gain and bandwidth are determined to be 7.8 dBi and ~40% for a microstrip fed antenna device while they are 7.6 dBi and 38% for a CPW fed device.

3.1 Introduction to aperture coupled patch antenna

The design of microstrip antennas at microwave and millimeter wave frequencies is closely related to the feeding technique. At these frequencies, there are several problems such as feed dimensions, soldering of probes associated with the classical feeding techniques, such as coaxial probe (Fig 1 (c)) or edge feeds (Fig. 1 (a) and (b)). These considerations are even

more important for wideband applications, which require thicker substrates. On the other hand, the aperture coupled feeding technique (Fig. 1 (d)) has intrinsic properties which make it an attractive feature for millimeter wave applications. Wide-band operation of this type of microstrip fed antenna has been demonstrated at microwave and millimetre wave frequencies using either single or stacked patch configurations. Although all of the coupling methods depicted in Fig.1 have been shown to give excellent bandwidth characteristics, the indirect methods (Fig. 1 (b) and (d)) give rise to a high back-radiation level. However this is only true for aperture coupling (Fig. 1 (d)) if the aperture is near resonance. The aperture coupled antenna can also be fed with a CPW feed and is ideal at millimeter wave frequencies in the 30 GHz range (Mestdagh et al, 2004). It has been found that these antennas can easily be impedance-matched by tuning the dimensions of the excitation aperture and adding a small tuning stub.

Numerous advantages accrue for an aperture coupled configuration including those summarized previously (Pozar, 1985). They are listed below:

(i)   The configuration is suited for monolithic phased arrays, where active devices can be integrated on, for example, a gallium arsenide substrate with the feed network, and the radiating elements located on an adjacent (low dielectric constant) substrate, and coupled to the feed network through apertures in the ground plane separating the two substrates. The use of two substrates avoids the deleterious effect of a high-dielectric-constant substrate on the bandwidth and scan performance of a printed antenna array.

(ii)   No radiation from the feed network can interfere with the main radiation pattern, as the ground plane separates the two mechanisms.

(iii)   No direct connection is made to the antenna elements, so problems such as large probe self reactances or wide microstripline (relative to patch size), which are critical at millimetre-wave frequencies, are avoided.

(iv)   Ideal for micromachined antennas. The fabrication of a directly coupled feed probe would involve many fabrication steps.

(v)   The aperture coupled feeding technique has intrinsic properties which make it an attractive feature for millimeter wave applications.

(vi)   Wide-band operation of this type of microstrip antenna has been demonstrated at microwave frequencies (1 - 10 GHz) using either single or stacked patch configurations.

(vii)   A simple aperture coupled antenna structure gives rise to a high back-radiation level if the aperture is near resonance - this problem is eliminated by adopting a stacked antenna configuration.

A simple configuration of an aperture coupled microstrip antenna is shown in Fig. 1(d). It consists of a radiating patch on one substrate coupled to a microstripline feed on another parallel substrate, through an aperture in the intervening ground plane. It should be noted that the aperture coupled microstrip antenna can be used for both linear and circular polarizations. It requires two co-located orthogonal apertures each one excited by a different feed line. Each aperture excites orthogonal linearly polarized resonances under the normally square patch. The polarisation of the radiation from the patch is then dependent on the relative magnitude and relative phase of the signals entering the independent feed lines.

Circular polarisation is obtained when the signals are equal in magnitude and in quadrature phase.

For conventionally fed patch antennas (Fig. 1 (a) to (c)) it is well known that to a very good approximation the patch resonant frequency is dictated largely by the size and shape of the patch. This is not the case in aperture coupled patches. The aperture also has a resonant frequency and the coupled resonators formed by the aperture and the patch resonates at a frequency determined by simple filter theory. In this section a micromachined microstrip fed aperture coupled antenna is studied in depth. We then examine micromachined aperture coupled stacked antenna devices fed both from microstrip and CPW line. Antenna performances are assessed for various antenna configurations. Reflection coefficient (S_{11}), VSWR, normalise radiation pattern, gain, directivity and efficiency parameters are presented for a range of design parameters with the results plotted as a function of frequency as necessary. The effects of these different design parameters on the antenna performance are discussed.

3.2 Micromachined aperture coupled patch antenna

The gain and bandwidth of an aperture coupled antenna device can be significantly improved by introducing a micromachined air gap below the patch (see Fig. 2). These aperture coupled antenna devices which are described in ensuing sections are produced on a microwave PCB substrate. The air gap is achieved by suspending the patch on a micromachined polymer ring structure. A brief introduction to an established analytical design method is provided in order to communicate the fundamental operating principle of an aperture coupled antenna of the type described here. However, an electromagnetic field full-wave simulation software package is employed to design the antenna devices and adjust them for optimum performance. The effects of the design parameters, such as substrate material, air gap thickness, polymer rim dimensions and conductor materials, on the antenna's electromagnetic behaviour are investigated. The performances of several micromachined antenna devices are compared. Microstrip fed, CPW fed, single and stacked antenna configurations are included.

An aperture coupled antenna structure can be effectively modelled by means of a range of well established techniques. These are:

- Transmission line model (TLM)
- Finite element model (FEM)
- Finite difference time domain technique (FDTD)
- Method of moments technique (MOM)

All of these techniques exist in commercial packages. The general purpose modelling package, ANSOFT HFSS, is based on the finite element method, while CST Microstripes is based on the transmission-line matrix (TLM) method in time domain form. The IE3D package employs the method of moments. All are suitable for the kind of micromachined antenna devices described in this chapter. The antennas presented here are all modelled and optimised in the ANSOFT HFSS design environment.

The micromachined cavity-backed patch antenna design, realized by means of a polymer spacer, is quite similar to the strip-slot-form-inverted patch (SSFIP) antenna (Zürcher, 1988).

Fig. 2 shows a schematic of the cross-sectional and the top views of this potentially high gain, wide band antenna device. A square polymer ring (SU8 rim) mimics a sealed air filled cavity between the substrate and the polyimide thin film. The cavity-backed aperture coupled configuration improves radiation efficiency and thereby the gain of the antenna device. The configuration also improves the bandwidth of the antenna device owing to the proximity of the resonances of the coupling aperture and the patch.

(a)

(b)
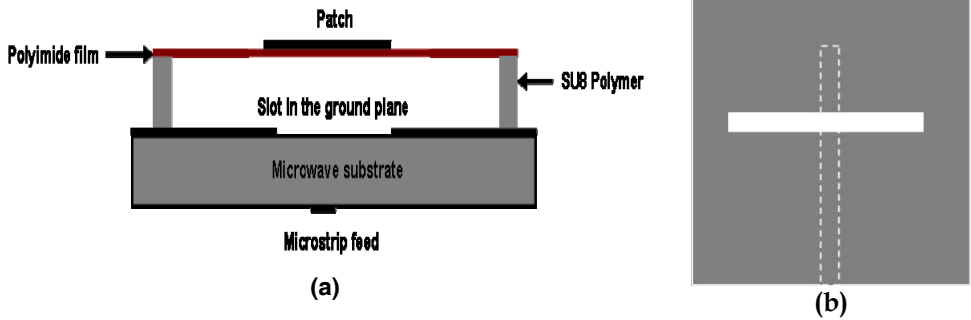
Fig. 2. Geometry of the (a) cross-sectional view and (b) top view of single patch micromachined aperture coupled antenna device.

The antenna centre frequency depends primarily on the dimensions of the resonant patch element and is given by

$$f_0 = \frac{c}{2L_{eff}\sqrt{\varepsilon_{eff}}}$$

(1)

where f_0 is the centre frequency of the antenna, c is the speed of light, L_{eff} is the effective length of the patch element and ε_{eff} is the effective dielectric constant. The device was designed to operate at around 12 GHz for ease of characterization using in-house measurement facilities. The design parameters such as the dimensions of the microstrip feed, aperture and top patch were optimized using the Ansoft HFSS electromagnetic simulation package with the aim of achieving high antenna radiation efficiency.

Parameter	Unit (mm)
Patch length	9.5
Patch width	11.5
Patch thickness	0.007
Slot length	11
Slot width	1.1
Microstrip length	18.84
Microstrip width	0.85
Inner length of cavity	19
Inner width of cavity	19
Cavity height	1.5
Thickness of polyimide film	0.125
Thickness of microwave PCB substrate	1.51

Table 1. Summary of the design parameters for the microstrip fed suspended patch antenna.

The height of the polymer rim was adjusted to obtain high gain. In the HFSS simulation, the lateral dimensions of the polymer rim were chosen to achieve high efficiency and compact size for the antenna structure. Table 1 summarizes the design parameters of the antenna device. A microwave PTFE material, AD300A, from Arlon MED (http://www.ctsind.com.sg/arlon.html), was used as the base substrate while a polyimide thin film (Du Pont) provided the supporting substrate for the suspended patch. The dielectric constant and loss tangent of the PTFE substrate are respectively 3 and 0.003. For the SU8 rim material they are 4.2 (Thorpe et al., 1998) and 0.042 (Lucyszyn, 2001) respectively, while the corresponding values for polyimide substrate material are 3.5 and 0.0026. In order to accurately model the performance of the antenna device, the parasitic effects of the SMA connector are simulated by introducing a short and wide rectangular extension at the input of the microstrip feed line. The dimensions (3mmx1.5mm) of the extension are chosen to match the length and diameter of the pin of the SMA connector. It has been observed that accommodating the effect of the connector is vital in order to accurately model the reflection characteristics of the suspended patch antenna. In the following sections the S_{11} parameters for estimating the bandwidth, normalised radiation pattern and radiation efficiency obtained from the HFSS designer environment are presented and discussed.

3.2.1 S parameters and study of the bandwidth
To assess the performance of the micromachined aperture coupled antenna device shown in Fig.2, the return loss (reflection coefficient in dB) and VSWR (voltage standing wave ratio) parameters are plotted for an optimum antenna, as in Fig. 3.
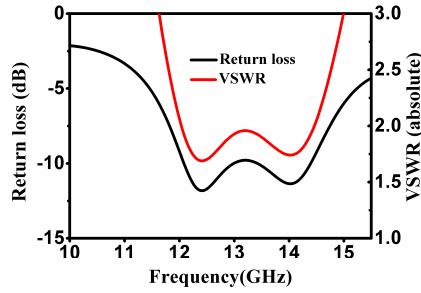
Fig. 3. Simulated return loss and the VSWR parameters for the optimized micromachined aperture coupled antenna device**.**

The return loss and VSWR parameters are plotted as a function of frequency from 10 to 15.5 GHz. It can be seen from the plot that the return loss is lower than -10 dB from around 12 GHz to 14.3 GHz and the VSWR is lower than a value of 2 in this range. Therefore the corresponding theoretical bandwidth of the antenna is 2.3 GHz or 17%.

3.2.2 3D antenna radiation pattern

The radiation pattern for the antenna device is obtained from the electromagnetic field solution and is plotted as a function of frequency from the far field plotter interface in Ansoft HFSS. Figure 4 shows the 2D and 3D far-field radiation patterns for the aperture coupled antenna device at 13.2 GHz near to the centre frequency of the operating band. The patterns show that there is high backward radiation and obvious side lobes in the E plane. But it will be shown in the later sections that the side lobes and back side radiation are reduced significantly using stacked patch configurations.
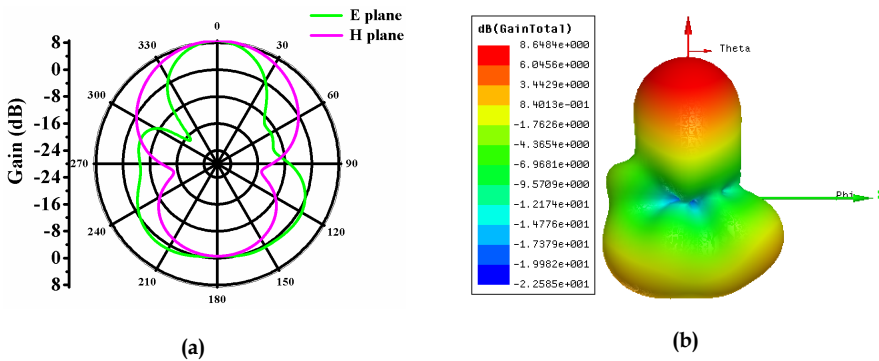
Fig. 4**.** 2D (a) and 3D (b) radiation patterns of a microstrip fed single patch antenna device at 13.2 GHz.

3.2.3 Directivity and gain

Fig.5 shows the variation of directivity, gain and efficiency for the microstrip fed aperture coupled antenna device with respect to frequency from 10 GHz to 15.5 GHz. The left 'y' axis gives the magnitude of antenna directivity and gain while the right 'y' axis gives the magnitude of absolute efficiency from 0.9 to 1. It can be easily seen from the plot the gain curve follows the directivity curve suggesting almost 100% radiation efficiency within the radiation bandwidth region. The gain varies from about 5 dBi at 10 GHz to around 8.3 dBi at 13 GHz and falls back below 6 dBi after 15.5 GHz. The excursions of directivity and gain within the radiation bandwidth (12 - 14.3 GHz) are below 1 dBi while it varies very significantly outside of this frequency range. The radiation efficiency varies from about 0.95 to close to 1 at 15.5 GHz. It should be noted that the radiation efficiency is a measure of total loss, including dielectric and conductor losses, within the micromachined antenna structure and it improves dramatically with the introduction of the micromachined air cavity.
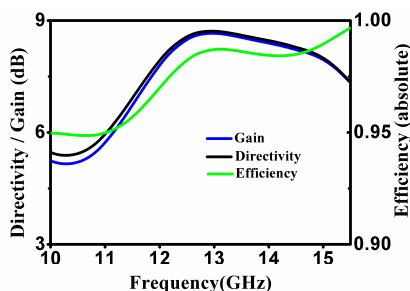
Fig. 5. Simulated gain, directivity and radiation efficiency for the optimized micromachined antenna device.

3.3 Stacked patch antenna devices

The design and modelling of the microstrip and CPW fed, stacked aperture coupled antenna devices is presented in the following sections. These stacked antenna devices are designed using a similar micromachined aperture coupled model as that presented in section 3.2. The stacked antenna configuration is exploited to further enhance the bandwidth and radiation pattern. The bandwidth is improved by utilizing the multiple closely resonant structures as well as through better impedance matching with the lower patch elements in the stacked configuration. The stacked patch elements are fabricated on polyimide substrates for microstrip fed devices and on LCP (liquid crystal polymer) film substrates for the CPW configurations, supported, in all cases, by the micromachined SU8 polymer spacers

3.3.1 Microstrip fed aperture coupled stacked patch antenna

3.3.1.1 Antenna Design

Stacked, cavity backed, aperture coupled antenna geometries have been modelled using Ansoft HFSS and optimized. Fig. 6(a) provides a cross sectional view of a microstrip fed stacked antenna of the type described in this section, while the top view of the coupling aperture on the lower microstrip substrate surface is depicted in Fig. 6(b). The arrangement consists of a double cladded microwave PTFE substrate, and two suspended patches to form

a stacked antenna device. The microstrip feed line on the bottom surface of the microwave substrate excites the lower patch through a rectangular coupling aperture in the ground plane which forms the top surface of the microwave substrate. The patch elements are printed on thin film (polyimide) substrates. They are supported by micromachined polymer spacers to reduce the antenna loss and hence to improve its gain. The cavities are sealed by the bonded polymer rings thus providing protection against the incursion of moisture and particles, which could contribute unwanted losses.

(a)   (b)
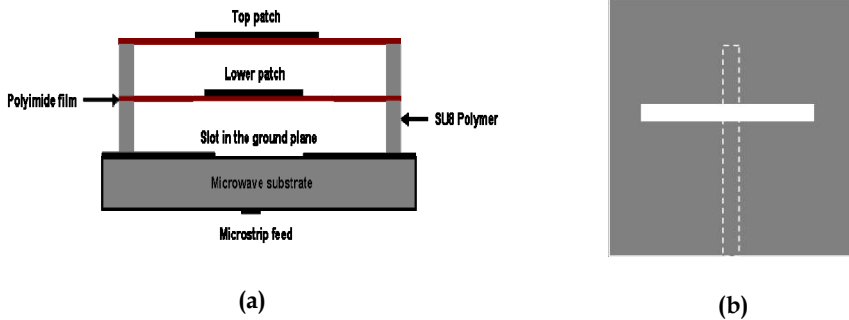
Fig. 6. (a) Schematic cross-sectional view of the sta   microstrip fed aperture coupled antenna using micromachined polymer spacers, (b) The top view of the aperture and feed line on the substrate surfaces.

Parameter	Microstrip fed device (mm)	CPW fed device (mm)
Lower patch length	5.8	9.4
Lower patch width	9	18
Upper patch length	9.6	12.6
Upper patch width	10.6	18
Patch thickness	0.009	0.009
Slot length	9.8	16.5
Slot width	0.98	1.6
Microstrip length	15.00	18.2
Microstrip width	0.82	1.4
Inner length of cavity	19	18.2
Inner width of cavity	19	18.2
Lower cavity height	2.2	1.0
Thickness of polymer rim	1.0	1.0
Upper cavity height	0.67	1.0

Table 2. Summary of the design parameters for the suspended stacked patch antennas

The stacked antenna has been designed using a similar approach to that described previously for the development of multi-layer stacked wideband antenna devices (Pavuluri et al., 2008, Wang and Pavuluri, 2008, Croq and Pozar, 1991). It is designed for operation at

X-band (8 – 12 GHz) with a 40% of bandwidth. The height of the air cavities formed by the micromachined polymer rims located between each stacked patch and the substrate, is adjusted for optimum bandwidth while at the same time maintaining a low profile for the overall antenna structure. For this microstrip fed device, the top patch and the aperture are dimensionally adjusted to ensure resonance at similar frequencies, while the dimensions of the microstrip line and the lower patch are tuned to secure the best possible impedance match. In the design process the initial dimensions of the patches are determined by the choice of the center frequency of operation. The width of the coupling aperture is conventionally set at about 10% of its length.

The heights of the air cavities as explained earlier are a trade-off between the desire for bandwidth and the limitations set by the fabrication process. The length of the feedline is then varied to obtain sufficient bandwidth. In order to obtain fixed band performance, the lengths of the top patch and the aperture are modified to tune the band of operation. Fine impedance tuning is achieved by adjusting the dimensions of the lower patch. The shape and dimensions of the SU8 spacer rings are chosen carefully to ensure good antenna performance. It is generally necessary to carry out several iterations of the above steps to obtain an acceptable antenna delivering high gain and wide bandwidth. The optimized design parameters are given in Table 2.

3.3.1.2 S parameters and study of the bandwidth

To determine the predicted performance of the optimally designed aperture coupled antenna, the return loss and VSWR are plotted as a function of frequency. The plot width at the -10 dB return loss locations as in Fig.7, quantifies the bandwidth. From the plot the return loss is lower than -10 dB from about 7.8 GHz to 12 GHz. The VSWR is lower than a value of 2 over this range. Therefore the theoretical bandwidth of this device is 4.2 GHz or 42%. This is an improvement of a factor of 2.5 over that of the single patch device.
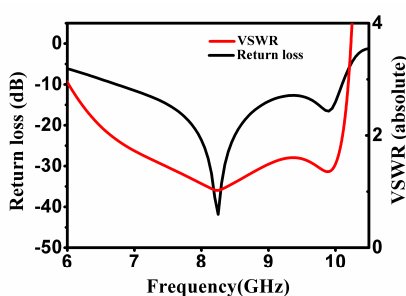

(a)
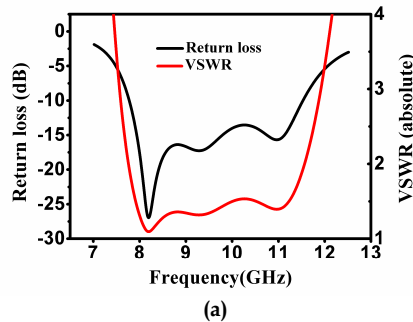
Fig. 7. Simulated insertion loss and VSWR parameters for the optimized micromachined microstrip fed stacked aperture coupled antenna.

3.3.1.3 3D antenna radiation pattern

The radiation pattern for the antenna device is obtained from the electromagnetic field solution and is plotted as a function of frequency using the far field plotter interface in Ansoft HFSS. Figure 8 shows the 2D and 3D far-field patterns for the aperture coupled

antenna device at 9.82 GHz near to the centre frequency (9.9 GHz) of the operating band. It can be seen that the back side radiation is reduced to around -4 dB for the microstrip fed stacked patch antenna configuration.

(a)   (b)
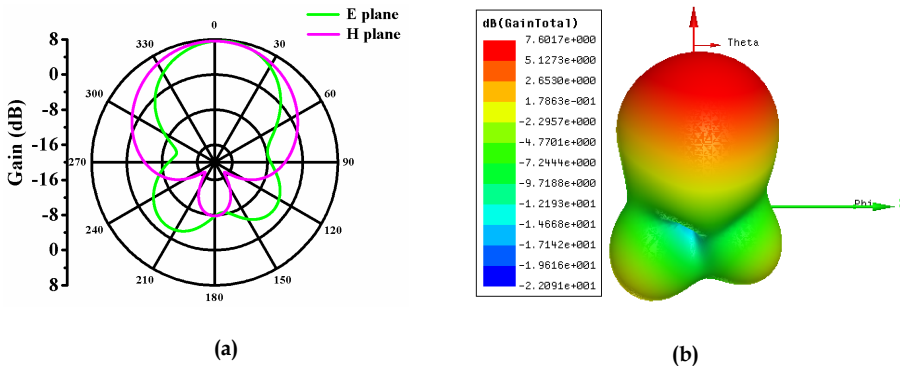
Fig. 8. 2D (a) and 3D (b) radiation patterns at 9.82 GHz.

3.3.1.4 Directivity and gain

Fig. 9 shows the dependence of directivity, gain and efficiency on frequency, in the range from 7 GHz to 12 GHz, for the microstrip fed stacked antenna. The results are not too dissimilar to those of the single patch antenna (Fig. 4), but with less gain and efficiency variation over the optimum operating band. The gain and directivity vary from about 6 dBi at 8 GHz to around 7.8 dBi at 9.7 GHz and falls back below 6 dBi after 12.5 GHz. The directivity and gain are constant to within 2 dBi over the previously defined -10dB bandwidth. Outside this bandwidth gain diminishes significantly. The radiation efficiency is greater than 0.95 from 8 GHz to 12 GHz.
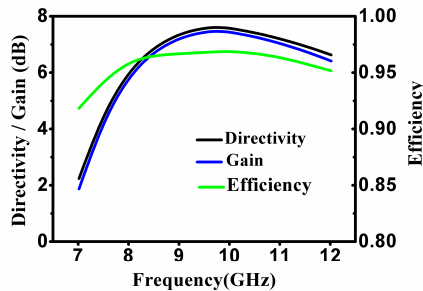
Fig. 9. Simulated gain, directivity and radiation efficiency of the optimized micromachined microstrip fed stacked aperture coupled antenna.

3.3.2 CPW-fed aperture coupled micromachined patch antenna

An antenna comprising three stacked patches fed from a coplanar waveguide (CPW) has also been studied. Fig. 10 show the schematics of the CPW fed stacked patch antenna. A single cladded PTFE substrate (Taconic TLY-3-0200-CH/CH) was used to support the CPW line and the coupling aperture. The stacked patches are suspended symmetrically above the

aperture using micromachined SU8 polymer rims. As the air gap between the base microwave substrate (Fig. 10(a)) and the lowest patch is smaller than that of the microstrip fed stacked device (Fig. 6(a)) for ease of fabrication, an additional patch is required to yield a similar bandwidth of ~40%. The upper pair of patches has the same dimensions and it is the electromagnetic coupling between them that further increases the bandwidth of the antenna. The coupling slot on the ground plane and the top two patches were designed to be in close resonance. The bottom patch and the λ/4 stub are optimized for a wide 'match' bandwidth. Tables 2 and 3 show the physical dimensions of the structure layers for the antenna device.

(a)   **(b)**
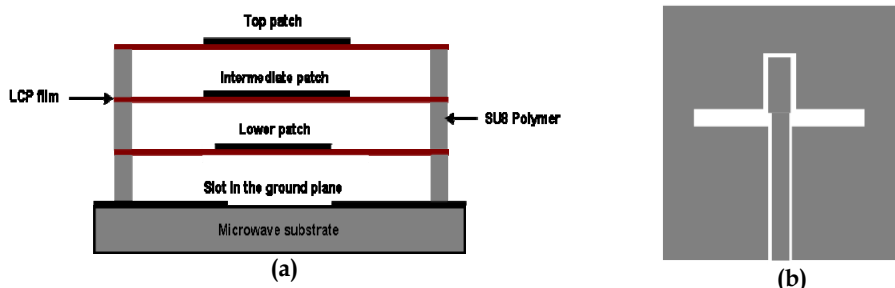
Fig. 10. (a) Cross-sectional view of the stacked CPW fed antenna using micromachined polymer spacers, (b) top view of the corresponding aperture and feed line on the substrate surface.

Substrate	Thickness (mm)	Dielectric constant	Loss tangent
Taconic PTFE substrate	0.5	2.2	0.0009
LCP film	0.10	3.2 (Thompson et al., 2004)	0.002

Table 3. The thickness and microwave property of Taconic PTFE and LCP substrates.

Fig. 11. Simulated insertion loss and VSWR parameters for the optimized micromachined CPW fed stacked aperture coupled antenna.

3.3.2.1 S parameters and study of the bandwidth

The return loss and VSWR parameters are plotted as a function of frequency from 6 to 10.5 GHz in Fig.11. It can be seen from the plot that the return loss is lower than -10 dB from

around 6.5 GHz to about 10.2 GHz, and the VSWR is lower than a value of 2 over this range. The corresponding theoretical bandwidth is 44%.

3.3.2.2 Radiation pattern and gain

As with earlier examples, the radiation pattern for the antenna device is obtained from the electromagnetic field solution. It is plotted at a given frequency using the far-field plotter interface in Ansoft HFSS. Fig. 12 shows the 2D and 3D far field patterns for the aperture coupled triple patch antenna device at 8.3 GHz. It can be seen that the back side radiation is reduced to around -7 dB in the E and H plane radiation characteristics. The backward radiation is significantly less than that of the microstrip fed devices. The reduction of the back radiation can be attributed to the effect of the CPW based feeding method. Thus the CPW feeding approach is recommended for minimal backward radiation.
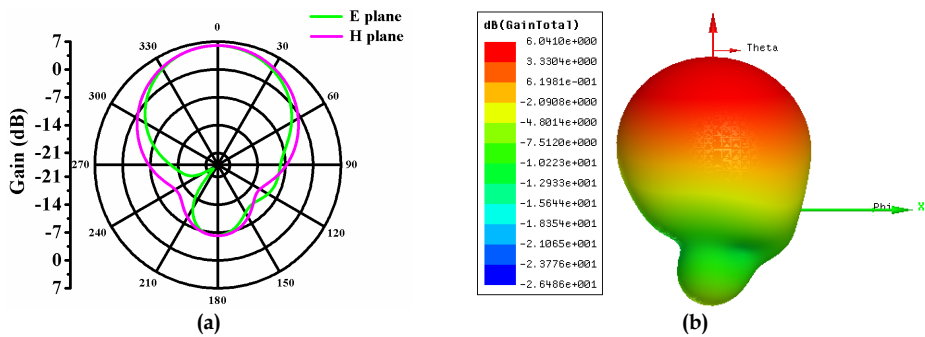
Fig. 12. 2D (a) and 3D (b) radiation patterns for the CPW fed stacked antenna at 8.3 GHz.

3.3.3 The effect of polymer rim design on the performance of the CPW fed stacked aperture coupled antenna

As stated previously the SU8 polymer has high dielectric loss at microwave frequencies with a loss tangent of 0.043. The effect of the dimensions of the polymer rim on antenna performance has been studied using the CPW fed antenna design. Two rim designs of 23mmx23mm and 12.5mmx18mm are used and the corresponding projected views of the antennas are shown in Fig. 13. The other design parameters remain the same as shown in Table 2. Fig. 14 shows the efficiency of the CPW fed stacked antenna device for the two polymer rim designs in the frequency range of 6.5 GHz to 10.5 GHz. As can be seen from Fig. 14 the efficiency of the antenna device for the larger rim is greater than 0.95 from 6.5 GHz to 10.5 GHz but it is less than 0.9 with the smaller polymer rim. It can also be seen that the roll-off rate of the antenna efficiency above 9.5 GHz is much faster for the smaller rim indicating rapid decrease of the antenna performance.

(a) **(b)**
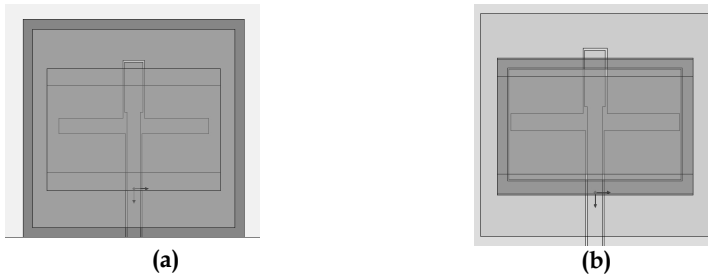
Fig. 13. A 3D view of the optimized micromachined CPW fed stacked aperture coupled antenna with two different polymer rim dimensions.
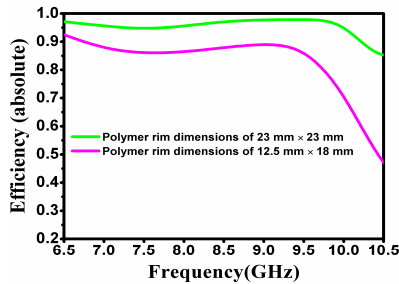
Fig. 14. Results of antenna efficiency of the CPW fed antenna for different SU8 polymer rim dimensions

3.3.4 Effect of the substrate loss

As the losses in the substrate layers in a stacked antenna affects its performance significantly, for comparison, FR4 and PTFE based stacked CPW antenna designs with similar dimensions as that of the CPW antenna shown in Fig. 10 are also designed and optimized for impedance matched performance. These two devices consist of 4 layers of FR4 or PTFE material with 3 three stacked patches. The dielectric constant and the loss tangent for the FR4 material are taken as 4.2 and 0.020 respectively (Aguilar et al., 1998). Fig. 15 shows the efficiency and gain as a function of frequency for the three CPW fed stacked antenna configurations. The rapid decrease of gain above 9 GHz of the FR4 based device is due to the increased insertion loss as the frequency is out of the band of operation. The gain of the antennas was obtained from the simulation results of the radiation characteristics. It can be seen that there is little difference between gain values of the stacked patch antenna based on the multilayer PTFE structure and the devices with suspended patch elements since the dielectric loss is minimal in both cases. However, the micromachined device has a larger bandwidth. The performance of the FR4 based multilayer antenna is much poorer due to the well known lossy behaviour of the FR4 material beyond the microwave frequency region. Table 4 gives a summary of the performance parameters. The micromachined antenna device with suspended patches showed the best bandwidth of about 38% close to that required for ultra-wide band applications.
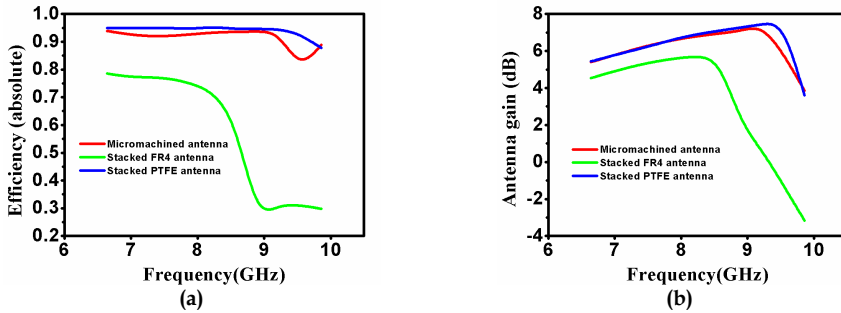
Fig. 15. (a) Results of antenna efficiency of the CPW fed stacked patch antenna for different substrate configurations, (b) the corresponding results for the gain of the devices.

CPW fed antenna design	Bandwidth (%)	Peak gain (dBi)	Efficiency (%)
Suspended patches on PTFE substrate	38.3	7.8	97.5
Stacked PTFE substrates	30.1	7.9	99
Stacked FR4 substrates	32.7	3.6	75

Table 4. Summary of performance parameters for different antenna configurations.

4. Fabrication and assembly

4.1 Overview of antenna fabrication and assembly

New fabrication and assembly methods have been developed to produce the antenna devices described in the previous sections. The antenna structures are fabricated in layers and then assembled using micromachined polymer spacers. Fig. 16 shows the schematic of the layers of a microstrip fed antenna with multiple patches to illustrate the low temperature assembly method. The feeding and coupling structures are produced on a PCB substrate using the conventional PCB technology or a microfabrication method for devices requiring high resolution metal line and aperture structures. The patches and the supporting micromachined polymer structures are fabricated on thin film polymer substrates. Polyimide and liquid crystal polymer thin films are suitable materials as the substrates for supporting the patches. After fabrication of the patch, the millimeter thick polymer structure for supporting the thin film substrate is fabricated using a surface micromachining method. An additional polymer spacer can be fabricated on the PCB substrate in order to produce a larger air gap between it and the polymer substrate about it. In this case the two spacers should be matched. The layers of the antenna are then aligned and bonded together to produce a stacked patch antenna device.
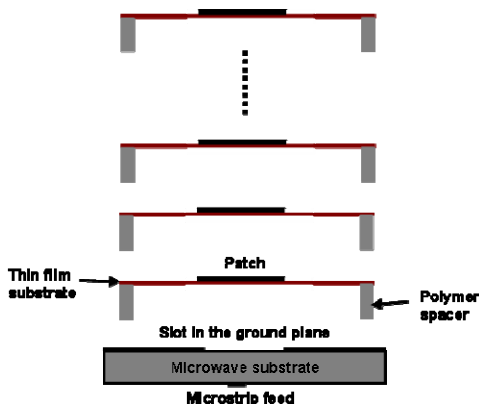
Fig. 16. Illustration of the assembly method for construction of a generalized micromachined antenna device.

4.2 Fabrication of antenna layers

All of the antenna layers are fabricated using microfabrication methods though some of the antenna structures such as the patches can be fabricated using the conventional PCB technology. A MEMS based fabrication process has been implemented to obtain precision metal patterns on the substrates with the aim of reducing metal losses at higher frequencies. In general, the fabrication process outlined in the following sections can be implemented with different substrate materials as long as the thermal stress due to the CTE (coefficient of thermal expansion) mismatch between the SU8 polymer and the substrates can be accommodated without the risk of the delamination of the SU8 structures from the substrates. In addition, this fabrication process can be used for constructing other passive devices such as high performance filters at millimetre wave and terahertz frequency regimes.

4.2.1 Substrate fabrication

Commercial microwave PCB substrates (Arlon MED300) are used as the base substrate for the microstrip fed antenna devices described in the previous section. The microstrip and the coupling aperture are fabricated using the copper cladding layers on the substrate. Fig. 17 shows the schematic of the fabrication steps. After cleaning and drying, both sides of the substrate are deposited with a photoresist layer (AZ9260, Clariant). The photoresist layers are deposited by spin coating and the thickness is about 10 µm. After deposition, the substrate wafer is left at room temperature for a few minutes to allow the air bubbles trapped in the resist film to propagate to the surface to eliminate the formation of defects in the photoresist layer after baking. In the photolithographic process the photoresist layer is exposed to UV light on a mask aligner using a high resolution polymer photomask to define a pattern for the microstrip line.
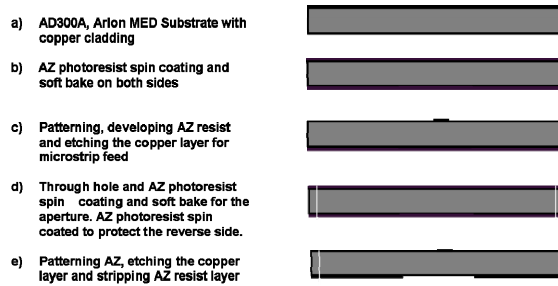
Fig. 17. Process flow of substrate fabrication for the microstrip fed antenna device.

The photoresist layer is then developed and the exposed copper is etched in a ferric chloride based etching bath. Careful control is necessary to ensure a well defined microstrip line. The photoresist layer is removed to complete the fabrication process for the microstrip line. For alignment of the microstrip to the aperture to be fabricated on the opposite side of the substrate in the subsequent fabrication process, precision holes are drilled at the corners of the substrate using a mechanical drilling method. Then both sides of the substrate are coated with a photoresist layer for fabrication of the coupling aperture and protection of the microstrip line. The same photoresist patterning and metal etching processes as for the microstrip are repeated to obtain the aperture. Fig. 18 shows the images of the fabricated microstrip line and aperture on the substrate. In the substrate fabrication process for the CPW fed device using PTFE substrates from Taconic only single sided metal patterning is required to fabricate the CPW structures.
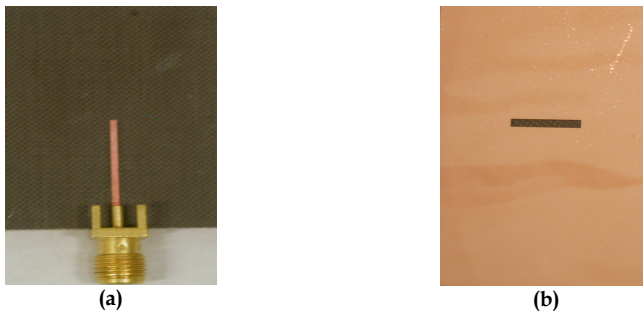
Fig. 18. Optical images of (a) a feeding microstrip and (b) a coupling aperture on a substrate after fabrication.

4.2.2 Fabrication of patches on thin film substrates

Polyimide and LCP (liquid crystal polymer) based films are used as the thin substrates for the suspended patches. The thicknesses of the polymer films are 125 µm and 100 µm respectively. These substrate materials are low-cost and are well known for the excellent dielectric, thermal and mechanical properties for microwave applications. As the film substrates do not have a copper cladding layer, a copper foil of 9 µm from Taconic is laminated on the substrate for patch fabrication. Alternatively the patch can be fabricated using electrodeposition of copper.
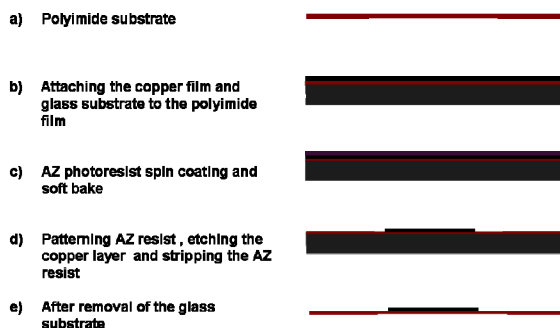
Fig. 19. Fabrication flow for the suspended patches.

In both cases the patch is defined by the photolithography method. As the thin film polymer substrate is flexible it is attached to a glass wafer for easy handling during patch fabrication. Fig. 19 illustrates the schematic of the process flow for patch fabrication on the polymer substrate following the lamination of the copper foil. A layer of photoresist is deposited on the substrate by spin coating and is then patterned by photolithography to define the patch structure. Following copper etching and photoresist stripping a patch is obtained on the thin film substrate. Several patches can be fabricated on a polymer film. The alignment marks are fabricated at the same time as for the patches.

4.2.3 Fabrication of micromachined spacers

The millimetre thick polymer spacer rims are fabricated on the thin film substrates after patch fabrication using the polymer surface machining method. Polymer micromachining is a process where high aspect ratio polymer structures are fabricated out of a thick polymer film on a substrate. We have used the photo-imageable SU8 epoxy polymer to produce the polymer structures on the substrate layers for constructing the antenna devices descried in the previous section. The SU8 polymer is a well known micromachining material for MEMS applications. It is a low cost material, easy to deposit in thick films by spin coating or spreading, low processing temperature (~110°C) and chemically inert. A wide range of film thicknesses from microns to millimetres can be obtained by modifying the composition of the chemical formulations for film deposition. The thick film and high aspect ratio capabilities have made the SU8 polymer an ideal material for the spacer rims for the antenna devices. The SU-100 polymer solution was prepared and used in order to fabricate millimeter thick polymer spacers. The composition of the solution is given in Table 6. Fig. 20 shows the process flow for fabrication of SU8 spacer rims on the substrate layers for antenna assembly. The required spacer thickness for each antenna layer is obtained by controlling the spin speed and time in the deposition of the SU8 layer.

Composition	Percentage
SU8 resin	72.9%
Gamma-butyrolactone liquid	21.1%
Photoinitiator	6%

Table 6. The composition of the SU8-100 solution.
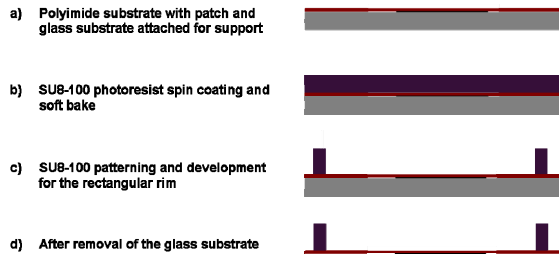
Fig. 20. Fabrication flow for the SU8 polymer rim.

After soft baking the film is exposed to the i-line (365 nm) of the UV source on a mask aligner using the photomask with the patterns of the spacer rims. The SU-8 layer is heated to 65^0C on a hotplate for post exposure bake. Then the SU8 layer is developed in the EC solvent to remove the unexposed SU8 in the photolithography process. Mild agitation provided by a magnetic stirrer can be used to reduce the development time and improve the quality of the resultant SU8 structures.

In order to create the large air gap between the microwave substrate and the suspended patch above it for the microstrip fed devices it is useful to fabricate two matching spacer rims one on each of the microwave and the film substrates. Fabrication of the millimetre thick SU8 spacer rims on the rigid Arlon MED substrate is more challenging than the thin film substrates and an optimized process for soft baking is necessary. Otherwise delamination of the SU8 spacer rims from the substrate occurs after the fabrication process due to the higher interfacial stress built up in the baking process prior to photolithography. The increased stress is caused by the larger CTE mismatch between the SU8 layer and the rigid microwave PCB substrate than the film substrate. However this problem can be solved by careful control of the temperature profile in the soft baking process, a smooth temperature profile in the heating and cooling processes is required (Patel et al., 2008). As a result millimetre thick SU8 polymer rims on microwave substrates are successfully fabricated.

4.3 Antenna assembly

The fabricated substrate layers are assembled using a low temperature polymer bonding process to form high performance antenna devices. This is a new method for creating antenna devices with suspended patch elements. Fig. 21 illustrates the process steps for assembly of a microstrip fed stacked patch antenna. A thin layer of SU8-100 solution is

applied to the top surface of the polymer rims as the bonding material for antenna assembly. The polyimide film supporting the lower patch is attached to the substrate by bonding of the SU8 rim on the polyimide film to the corresponding one on the base substrate. Then the polyimide film with the top patch was attached to the polyimide substrate of the lower patch by bonding the SU8 rim to the polyimide film. Precision alignment marks on the polyimide films and the base substrate were used to facilitate the accurate alignment of the layers. The bonding process can be carried out on a hotplate or on a flip chip bonder.
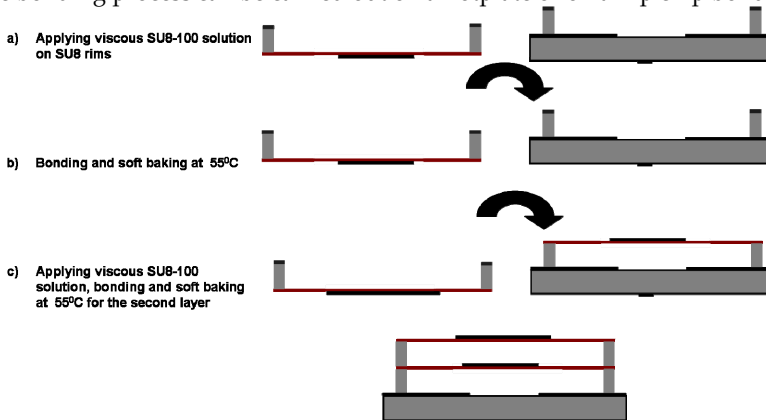
Fig. 21. Process flow for the antenna assembly.

The polymer rim on the lower surface of the polyimide substrate is aligned and attached to the rim on the microwave substrate. The assembly is heated to 55°C and maintained at this temperature for 10 minutes for bonding the two polymer rims together. The thin film substrate with the top patch is then assembled on top of the substrate layer with the lower patch using the same bonding approach. After the assembly process a stacked patch antenna device was obtained. For RF characterization to evaluate the performance of the device, an SMA connector was attached to microstrip and ground plane on the microwave substrate. CPW fed antennas are assembled using the same method Fig. 22 shows the images of the assembled antenna devices. For the CPW fed device, the patches are on the bottom surface of the LCP substrates. The most critical parameters determining the quality of the bonding are the thickness uniformity of the SU-8 rims, the amount of SU8-100 solution applied to the surface of the SU8 rims and the bonding pressure and temperature. SU8 rims with good surface uniformity can be obtained using an optimized fabrication process. The amount of liquid SU8-100 for bonding is determined by minimizing void formation and at the same time reducing the overflow of the material during bonding. A good bonding temperature is 65^0C which is just above the glass transition temperature (64^0C) of the SU8 material but not high enough to cause significant overflow as the viscosity changes rapidly around the glass temperature (Tuomikoski and Franssila, 2005).

(a)   (b)

Fig. 22. Optical images of the assembled stacked antennas (a) microstrip fed and (b) CPW fed device.

5. RF Measurements

In the following section, the RF measurements for S_{11} parameters and the radiation patterns along with the results of gain measurements are presented and compared with the Ansoft HFSS simulation results. The S_{11} parameters are obtained from a HP 8510 network analyzer after calibrating the coaxial cables. Radiation patterns are obtained by measuring the S_{21} parameters between the AUT (Antenna Under Test) and a reference 20 dBi standard gain horn in X band. The three antenna gain measurement technique was used for the micromachined aperture coupled single patch antenna device while a gain comparison technique was used for the stacked antenna configurations. The antenna measurements were conducted in an 8m×3m×3m anechoic chamber with a separation distance of 1.5 m between the reference and the AUT.

5.1 Results of reflection measurements

Fig. 23 shows the measured and simulated results of S_{11} for the single patch aperture coupled antenna device. The resonance at 14.2 GHz is that of the patch while the one at 12.1 GHz is associated with the slot. The -10dB bandwidth of the device was determined to be 19% from the measured results, a factor of 2 larger than that of the probe fed, post supported device (Pan et al, 2006).
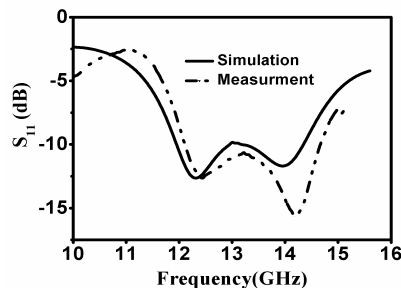
Fig. 23. Simulation and measurement results of the reflection characteristic of the high gain aperture coupled patch antenna.
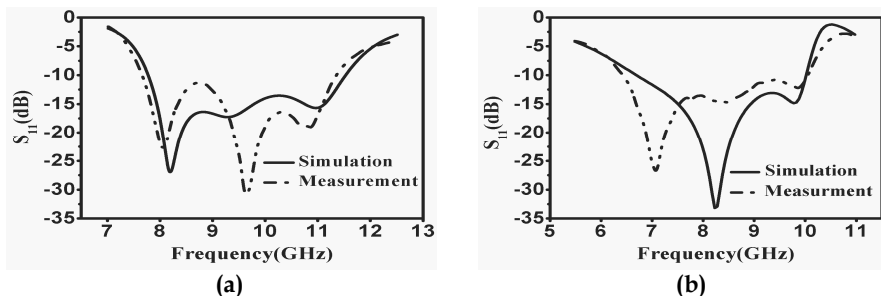
Fig. 24. Simulation and measurement results of the reflection characteristics for the high gain wideband stacked antennas, (a) microstrip fed device and (b) CPW design.

Fig. 24 shows the measured and simulated results of the reflection characteristic (S_{11}) of the microstrip fed and CPW fed stacked antenna devices. The -10dB bandwidth for the microstrip fed stacked patch antenna device is determined to be 3.9 GHz or 40% from the measured results. The corresponding result for the CPW fed stacked patch antenna is 3.25 GHz or 38%. It can be seen that the measured results agree well with that of the simulation apart from the downward frequency shift. This shift, particularly in the vicinity of the main resonance, can be attributed to the effect of the SMA connector overlay which was not included in the simulation.

5.2 Results of radiation measurements

Fig. 25 shows the gain-factor results for the micromachined single patch antenna device measured at 12.1 GHz. The gain of the antenna is about 8.3 dBi. Figs. 26 and 27 show the measurement and simulation results of the normalized radiation pattern for the stacked antenna devices. The maximum gain of the microstrip fed antenna devices is about 7.8 dBi at ~10.5 GHz while the maximum gain of CPW fed antenna device is about 7.6 dBi at ~10 GHz. The results show good agreement between measurement and simulation.
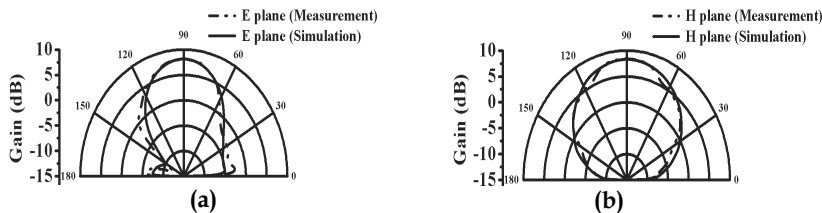
Fig. 25. Simulated and measured E-plane (a) and H-plane (b) radiation patterns for the microstrip fed single patch antenna device at 12.1 GHz.
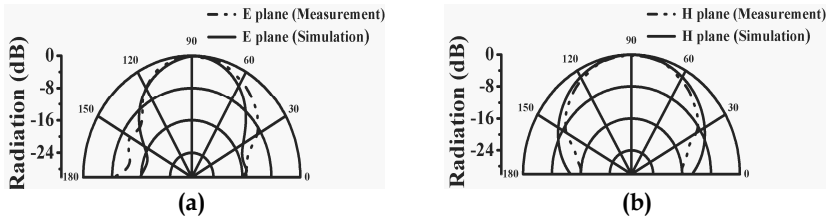
Fig. 26 Simulated and measured E-plane (a) and H-plane (b) radiation patterns for the microstrip fed antenna device at 10.5 GHz.
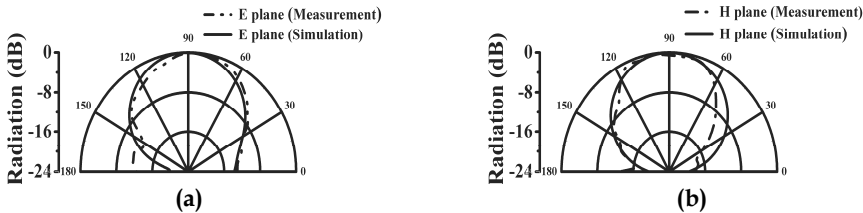
Fig. 27. Simulated and measured E-plane (a) and H-plane (b) radiation patterns for the CPW fed antenna device at 10 GHz.

6. Conclusions

Micromachined aperture coupled patch antenna devices have been developed using polymer micromachining and micro-assembly methods. The antennas are fabricated in layers and then assembled using micromachined polymer spacers and a low temperature polymer bonding method. By creating air cavities between the radiating patches and the substrates using polymer rims, the efficiency, gain and bandwidth of the devices have been improved significantly over conventional microstrip patch antennas. The potential radiation efficiency of the devices is as high as 98% based on the simulation results. The wideband operation is achieved by employing several controls. These are: optimally tuning resonances associated with the antenna device; through the use of stacked patches as impedance transformers by optimizing the dimensions of the lower patches; by adjusting the tuning stub and the input aperture dimensions. This polymer based stacked antenna process has a dual function of supporting the air suspended radiating patches and for antenna module packaging. The new fabrication method provides an alternative low cost packaging approach as compared to conventional LTCC and PCB technology. The suspended patch antennas are lightweight and can be fabricated easily using polymer micromachining and bonding approaches. Similar devices can be designed for other frequency bands to meet the requirements of potential applications in the emerging field of UWB networks, WLAN networks at 60 GHz and automotive collision radar. The fabrication method could be developed into an effective packaging technology for millimetre wave and terahertz based sensor and communication systems. Antenna array devices can also be fabricated for high gain wideband applications.

7. References

AGUILAR, J. R., BEADLE, M., THOMPSON, P. T. & SHELLEY, M. W. (1998). The microwave and RF characteristics of FR4 substrates. *Proceedings of the IEE Colloquium on Low Cost Antenna Technology (Ref. No. 1998/206),* pp. 2/1-2/6, London, 24 Feb. 1998.

Arlon materials for electronics (MED), orlando, florida, ([Online]. Available: http://www.ctsind.com.sg/arlon.html), Data sheet for AD 300A PTFE. (woven fiberglass/ micro-ceramic filled laminate for RF & microwave printed circuit boards).

BAHL, I. J. & BHARTIA, P. (1980). *Microstrip Antennas*, Artech House, ISBN 978-0890060988, Dedham, Massachusetts.

BARAS T. & JACOB, A. F. (2008). Manufacturing reliability of LTCC millimeter-wave passive components. *IEEE Transactions on Microwave Theory and Techniques,* Vol. 56, No. 11, (Nov. 2008), pp. 2574-2581.

BO, P., YOON, Y., PAPAPOLYMEROU, J., TENTZERIS, M. M. & ALLEN, M. G. (2005). A high performance surface-micromachined elevated patch antenna. *Proceedings of IEEE Antennas and Propagation Society International Symposium, (2005),* Vol. 1B, pp. 397-400, Washington DC, 3-8 July 2005.

CHEN, N. Z. N. (2008). Wideband microstrip antennas with sandwich substrate. *IET Microwaves, Antennas and Propagation,* Vol. 2**,** No. 6, (Sep. 2008), pp. 538-546.

CROQ, F. & POZAR, D. M. (1991). Millimeter-wave design of wide-band aperture-coupled stacked microstrip antennas. *IEEE Transactions on Antennas and Propagation,* Vol. 39, No. 12, Part 1, (Dec. 1991), pp. 1770-1776.

DU PONT INC., Summary of Properties for Kapton Polyimide. [Online]. Available: http://www2.dupont.com/Kapton/en_US/index.html.

GAUTHIER, G. P., COURTAY, A. & REBEIZ, G. M. (1997). Microstrip antennas on synthesized low dielectric-constant substrates. *IEEE Transactions on Antennas and Propagation,* Vol. 45, No. 8, (Aug. 1997), pp. 1310-1314.

HOU, F., ZHU, J., YU, Y.-W. & WU, J. (2008). Millimeter-wave design of aperture-coupled micromachined patch antennas. *Proceedings of the International Conference on Microwave and Millimeter Wave Technology* (ICMMT 2008), pp. 1401-1404, Nanjing, China, 21-24 Apr. 2008.

HUANG, C. U., CHEN, I. Y., CHEN, H. J. H., JOU, C. F. & HUANG, S. R. S. (2005). Compact 3D-MEMS-meander monopole antenna. *Electronics Letters,* Vol. 41, No. 21, (Oct. 2005), pp. 1149-1151.

JAMES, J. R. & HALL, P. S. (1989). *Handbook of microstrip antennas*, IET, ISBN 978-0863411502, London.

KOUL, S. K. (2007). Design and development of passive and active RF components using MEMS technology. *Proceedings of International Workshop on Physics of Semiconductor Devices* (IWPSD 2007), pp. 670-675, Bombay, India, 16-20 Dec. 2007.

KRATZ, H. & STENMARK, L. (2005). Micromachined S-band patch antenna with reduced dielectric constant. *Proceedings of the 13th International Conference on Transducers, Solid-State Sensors, Actuators and Microsystems,* pp. 1088-1091, Seoul, Korea, 5-9 June 2005.

LEUNG, L. L. W., CHEN, K. J., HUO, X. & CHAN, P. C. H. (2002). On-Chip microwave filters on standard silicon substrate incorporating a low-k BCB dielectric layer. *Proceedings of the 32nd European Microwave Conference,* pp. 1-4, Milan, Italy, Oct. 2002.

LUCYSZYN, S. (2001). Comment:Terahertz time-domain spectroscopy of films fabricated from SU-8. *Electronics Letters,* Vol. 37, No. 20, (Sep. 2001), pp. 1267-1267.

MENDES, P. M., POLYAKOV, A., BARTEK, M., BURGHARTZ, J. N. & CORREIA, J. H. (2006). Integrated chip-size antennas for wireless microsystems: Fabrication and design considerations. *Sensors and Actuators A: Physical,* Vol. 55, No. 2, (2006), pp. 217-222.

MESTDAGH, S., WALTER DE, R. & VANDENBOSCH, G. A. E. (2004) CPW-fed stacked microstrip antennas. *IEEE Transactions on Antennas and Propagation,* Vol. 52, No. 1, (Jan. 2004), pp. 74-83.

OJEFORS, E., KRATZ, H., GRENIER, K., PLANA, R. & RYDBERG, A. (2006). Micromachined loop antennas on low resistivity silicon substrates. *IEEE Transactions on Antennas and Propagation,* Vol. 55, No. 12, (Dec. 2006), pp. 3593-3601.

PAN, B., YOON, Y. K., PONCHAK, G. E., ALLEN, M. G., PAPAPOLYMEROU, J. & TENTZERIS, M. M. (2006). Analysis and characterization of a high-performance Ka-band surface micromachined elevated patch antenna. *IEEE Antennas and Wireless Propagation Letters,* Vol. 5, No. 1, (Dec. 2006), pp. 511-514.

PAPAPOLYMEROU, I., FRANKLIN DRAYTON, R. & KATEHI, L. P. B. (1998). Micromachined patch antennas. *IEEE Transactions on Antennas and Propagation,* Vol. 46, No. 2, (Feb. 1998), pp. 275-283.

PATEL, J. N., KAMINSKA, B., GRAY, B. L. & GATES, B. D. (2008). PDMS as a sacrificial substrate for SU-8-based biomedical and microfluidic applications. *Journal of Micromechanics and Microengineering,* Vol. 18, No. 9, (Sep. 2008), pp. 095028/1-095028/11.

PAVULURI, S. K., WANG, C. H. & SANGSTER, A. J. (2008). A high-performance aperture-coupled patch antenna supported by a micromachined polymer ring. *IEEE Antennas and Wireless Propagation Letters,* Vol. 7, (2008), pp. 283-286.

POZAR, D. M. (1985) Microstrip antenna aperture-coupled to a microstripline. *Electronics letters,* Vol. 21, No. 2, (Jan. 1985), pp. 49-50.

POZAR, D. M. (1992) Microstrip antennas. *Proceedings of the IEEE,* Vol. 80, No. 1, (Jan. 1992), pp. 79-91.

RYO-JI, T. & KUROKI, F. (2007). High permittivity tape transmission line embedded in low dielectric support at millimeter-wave frequencies. *Proceedings of the Microwave Conference,* (KJMW 2007), pp. 189-192, Okinava, Japan, 15-16 Nov. 2007.

SCHULER, K., VENOT, Y. & WIESBACK, W. (2003). Innovative material modulation for multilayer LTCC antenna design at 76.5 GHz in radar and communication applications. *Proceedings of the 33rd European Microwave Conference,* Vol. 2, pp. 707-710, Munich, Germany, 7-9 Oct. 2003.

TACONIC, SPECIALIZING IN COATING OF GLASS FABRICS WITH PTFE ([Online]. Available: http://www.taconic-add.com/pdf/tly.pdf) Data sheet for TLY RF & Microwave Laminate. *Woven fiber glass fabric coated with PTFE interleaved with thin sheets of pure PTFE.*

THOMPSON, D. C., TANTOT, O., JALLAGEAS, H., PONCHAK, G. E., TENTZERIS, M. M. & PAPAPOLYMEROU, J. (2004). Characterization of liquid crystal polymer (LCP) material and transmission lines on LCP substrates from 30 to 110 GHz. *IEEE Transactions on Microwave Theory and Techniques,* Vol. 52, No. 4, (April 2004), pp. 1343-1352.

THORPE, J. R., STEENSON, D. P. & MILES, R. E. (1998). High frequency transmission line using micromachined polymer dielectric. *Electronics Letters,* Vol. 34, No. 12, (June 1998), pp. 1237-1238.

TONG, K. F., AU, T. M., LUK, K. M. & LEE, K. F. (1995). Wideband two-layer five-patch microstrip antennas. *Digest of the International Symposium on Antennas and Propagation Society*, Vol. 4, pp. 2112-2115, California, 18-23 June 1995.

TUOMIKOSKI, S. & FRANSSILA, S. (2005). Free-standing SU-8 microfluidic chips by adhesive bonding and release etching. *Sensors and Actuators A: Physical,* Vol. 120, No. 2, (Feb. 2005), pp. 408-415.

WANG, C. H. & PAVULURI, S. K. (2008). Fabrication and assembly of high gain MEMS antennas for wireless communications. *Proceedings of the 58th Electronic Components and Technology Conference* (ECTC 2008), pp. 1941-1945, Florida, 26-29 May 2008.

WANG, J., CAI, Y., ATIVANICHAYAPHONG, T., CHIAO, M. & CHIAO, J. C. (2005). Fabrication techniques and RF performances of transmission lines on polymer substrates. *SPIE Proceedings of Conference on Microelectronics: Design, Technology, and Packaging II*, pp. 60351M.1-60351M.9), Brisbane, Australia, Dec. 2005.

WI, S. H., KIM, J. S., KANG, N. K., KIM, J. C., YANG, H. G., KIM, Y. S. & YOOK, J. G. (2007). Package-level integrated LTCC antenna for RF package application. *IEEE Transactions on Advanced Packaging*, Vol. 30, No. 1, (Feb. 2007), pp. 132-141.

WOLFF, I. (2007). Design and technology of microwave and millimeter wave LTCC circuits and systems. *Proceedings of the International Symposium on Signals, Systems and Electronics,* pp. 505-512, Montreal, Canada, 30 July – 2 Aug. 2007.

ZÜRCHER, J.-F. (1988). The SSFIP: A global concept for high performance broadband planar antennas. *Electronics letters,* Vol. 24, No. 23, (Nov. 1998), pp. 1433-1435.

Microstrip Antennas for Mobile Wireless Communication Systems

Hala Elsadek
Electronics Research Institute, Microstrip Department Cairo
Egypt

1. Introduction

1.1 Introduction to mobile communications

In the last few years, the trend of the mobile phone technology has been dramatically decreased the weight and size. Due to enhancement in this trend, the antennas used for mobile hand held devices have to be small, light-weighted, low profile, and have an omni-directional radiation pattern in the horizontal plane. However, still there are challenges in the antenna's performance during interaction with the user's head and hand. The movement of the user during usage of the mobile hand held device often lead to gain, radiation pattern and input impedance change. Therefore, antennas used in hand-held transceivers for personal communications have been recognized as crucial elements that can either improve or limit system performance. This is particularly true in terms of bandwidth and efficiency. Therefore, to carefully design a handset with superior performance, engineers need to give attention to the design of the antenna systems of the mobile transceiver.

1.2 Introduction to antenna systems for mobile wireless communications

Monopole antennas have been the best whelming choice for use in various automobiles and mobile equipments. This is basically due to its simplicity in design and having the characteristics that are essential for mobile communication. For antenna designers, there is always unprecedented demand to create compact or even electrically small antennas that are compatible with the modem technology, which will operate on a small handheld ground plane, and satisfy the performance specifications, particularly with respect to bandwidth and efficiency.

Typical antenna elements used so far for small mobile terminals are the monopole (MP) (M. D. Yacoub,1993; J. R. James & P. S. Hall, 1989), the dipole, the normal-mode helix (K. Fujimoto, 1998), the planar inverted-F(PIFA) (C. P. Huang, 1999), the microstrip patch (MSA) (J. R. James and P. S. Hall, 1989), and the meander line. Recently, there is a growing tendency in recent mobile phones to employ only a built-in antenna, instead of using a monopole that sticks out of the equipment case (C. P. Huang, 1999; D. M. Pozar and D.H. Schaubert, 1995) so, the use of a chip antenna which is a very small, ceramic encapsulated antenna has become popular.

2. Mobile Antenna Requirements and Challenges

The explosive growth in the demand for wireless communication and information transfer using handsets and personal communication (PCS) devices has created the need for major advancements of antenna design as a fundamental part of any wireless systems. In most cases, the handheld antenna is placed on a small plastic/conducting box that is in close proximity to biological tissue of user body and also the surrounding crowded electromagnetic waves. At the same time, the system must radiate low power and provide reliable communication of voice and possibly data (D. M. Pozar & D.H. Schaubert, 1995) Added to the operational requirements, the user and service providers demand wireless units with antennas that are small and compact, cost effective for manufacturability, low profile and easy to integrate with the wireless communication system. The antenna designer must also consider the following electrical characteristics of the antenna which include (W. Stuzman & G.Theile, 1998; C. A. Balanis, 1997).

 1-antenna tuning (operating frequency);
 2- VSWR and Return loss (input impedance);
 3- Bandwidth;
 4-Gain and directivity;
5- Radiation pattern;
6- Diversity;
But significant new variables are introduced as:
1-The size of the chassis (expressed as a function of wavelengths);
2- SAR (Specific absorption rate) of the antenna.
These design considerations have led antenna designers to consider a wide variety of antenna structures to meet the often conflicting needs for wireless systems as: operation frequency, return loss (VSWR), bandwidth, gain and directivity, location and orientation of the antenna, radiation pattern and diversity.

3. Microstrip Antennas

One type of antennas that fulfills most of the wireless systems requirements are the *microstrip antennas*. These antennas are widely used on base stations as well as handsets. Microstrip antennas have a variety of configurations and have been the topic of what is currently the most active field in antenna research and development. The microstrip patch antennas (J. R. James & P. S. Hall, 1989) has increasingly wide range of applications in wireless communication systems as mobile and satellite communication systems due to their great advantages. Simply, microstrip structure consists of a thin sheet of low-loss insulating material called the dielectric substrate. It is completely covered with a metal on one side, called the ground plane, and partly metalized on the other side, where the circuit or antenna patterns are printed. Components can be included in the circuit either by implanting lumped components or by realizing them directly within the circuit as shown in Figure 1.
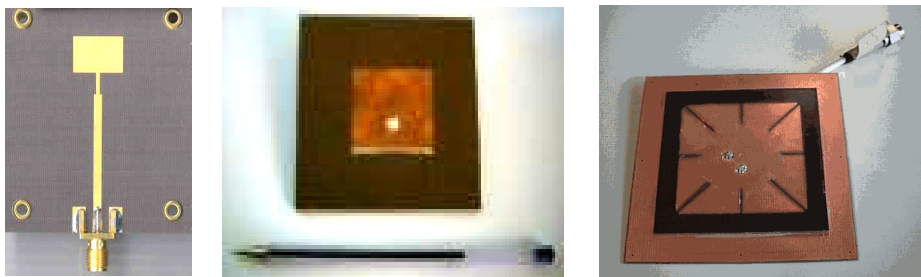
Fig. 1. Microstrip patch antennas

Microstrip antenna offers certain advantages as:

1. *Easy to handle and inexpensive*: as it is small and light
2. *Low power-handling capability* of printed circuits
3. *Topological Considerations*: integral with other circuit devices, hence, improving system reliability.
4. *Reduction in Size*: printed circuits are thin and thus require less volume than their waveguide or coaxial line counterparts.
5. *Large-Scale Fabrication*: once the designer has developed the basic circuit pattern, realized a circuit, and tested it successfully, additional copies can be produced rapidly and consistently in large mass production.

On the other hand, it suffers from some disadvantages as:

1. Narrow bandwidth.
2. Low gain.
3. Large ohmic loss in large feed network.
4. Radiation from feeds contributes to the radiation pattern.
5. Excitation of surface waves.
6. Low polarization purity.

3.1 Applications

Most of the rapid advances in microstrip antennas and arrays took place in the 1980s. Firstly, these were driven by defense and space applications. Then this technology is growing rapidly in the commercial sector. Specifications for defense and space application antennas typically emphasize maximum performance with little constraint on cost. On the other hand, commercial applications demand low cost components, often at the expense of reduced electrical performance. Thus, microstrip antennas for commercial systems require low-cost materials, simple and inexpensive fabrication techniques. Some of the commercial systems that presently use microstrip antennas are listed in the Table 1:

Application	Frequency
Global Positioning Satellite	1575 MHz and 1227 MHz
GSM	890-915 MHz and 935-960 MHz
Wireless Local Area Networks	2.40-2.48 GHz and 5.4 GHz
Cellular Video	28 GHz
Direct Broadcast Satellite	11.7-12.5 GHz
Automatic Toll Collection	905 MHz and 5-6 GHz
Collision Avoidance Radar	60 GHz, 77 GHz, and 94 GHz
Wide Area Computer Networks	60 GHz

Table 1. Microstrip Antenna Applications

3.2 Design Techniques

Another area of research, which is closely linked to advances in antenna development, is the field of *computational electromagnetics*. As antennas become more complex, the use of simple analytical modeling techniques (such as Transmission line model and cavity model) is not sufficient anymore. The use of full-wave modeling techniques, has therefore become inevitable. The most commonly used full-wave modeling techniques are the Method of Moment (MoM), the Finite Element method (FEM) and the *Finite Difference Time Domain (FDTD)*.

3.3 Microstrip antenna structure

A microstrip structure is illustrated in Figure 2.
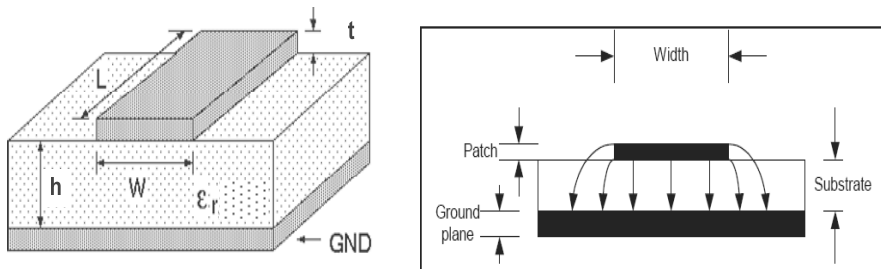
Fig. 2. Microstrip construction

3.3.1 Waves in microstrip structure

Depending on the direction toward which waves are transmitted, waves in microstrip fall within four distinct categories (W. Stuzman & G. Theile, 1998):

(1) Space waves: Waves transmitted upward, with elevation angles between 0 and $\pi/2$, move toward free open space, thus radiated, with field amplitudes decreasing with distance from antenna. In transmission lines and circuits, they are undesirable as they produce spurious leakage, however they contribute directly to antenna radiation efficiency as in figure 3 (a).

(2) Guided waves: they are trapped between the upper and lower conductors. They provide the normal operation for transmission lines and circuits but it is not favorable for patch antennas as in figure 3 (b).

(3) Leaky waves: they are directed more sharply downward, with θ lying between $\pi - \arcsin(1/\sqrt{\varepsilon_r})$ and $\pi/2$. These waves contribute to radiation thus they are useful for antennas as in figure 3 (c).

(4) Surface waves: the waves transmitted slightly downward, having elevation angles between $\pi/2$ and $\pi - \arcsin(1/\sqrt{\varepsilon_r})$, experience total reflection between the two conducting plates. The fields remain mostly trapped within the dielectric, decaying exponentially above the interface as shown in Figure 3 (d). Surface waves take up some part of the signal's energy thus reduces the antenna efficiency.

Space (radiated) Waves

(a)

Upper Conductor

Guided Waves

(b)

z

Leaky Waves

(c)

$\pi - \arcsin(1/\sqrt{\varepsilon_r}) < \theta < \pi$

z

Total Reflection

Surface Waves

(d)

Edge Diffraction

$\pi/2 < \theta < \pi - \arcsin(1/\sqrt{\varepsilon_r})$
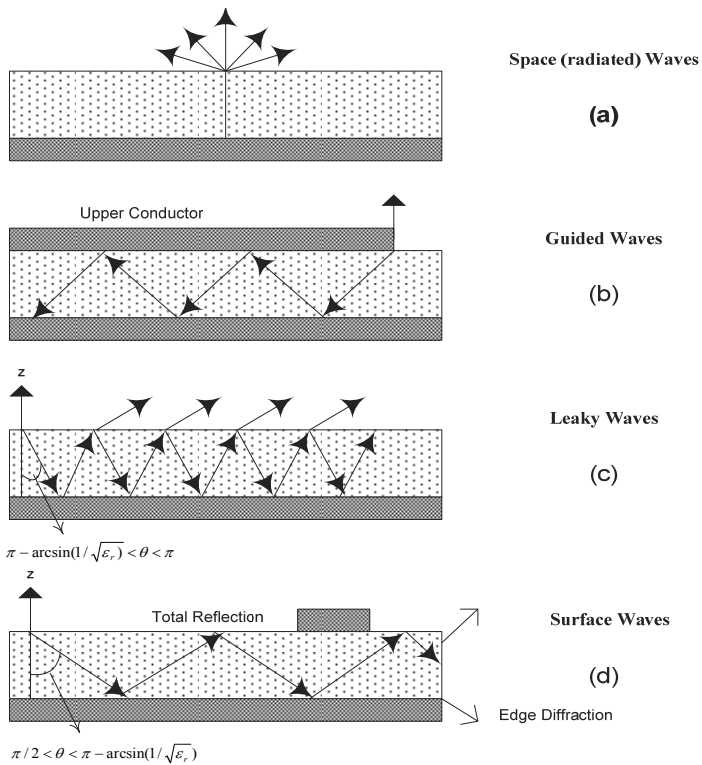
Fig. 3. Waves in microstrip

3.4 Microstrip array

Microstrip patch antenna is important as single radiating element, but the major advantages are realized in applications that require moderate size array as shown in Fig 4. When discrete radiators are combined to form an array, some characteristics can be achieved or enhanced; such as high gain and beam scanning.

The microstrip antenna arrays can be classified with different criteria. The elements of the array can be distributed to form linear, planar or volume array. Several ways of feeding can be used to obtain certain characteristics. The feed network of phased arrays or SISO arrays (Single Input Single Output) can be categorized into parallel and series feed. The parallel (corporate) feed has single input port and multiple feed lines. If the array has multiple inputs and multiple output ports so it is called Multi-element-array (MEA) or MIMO arrays (D. M. Pozar & D. H. Schaubert, 1995).

(a)   (b)   (c)
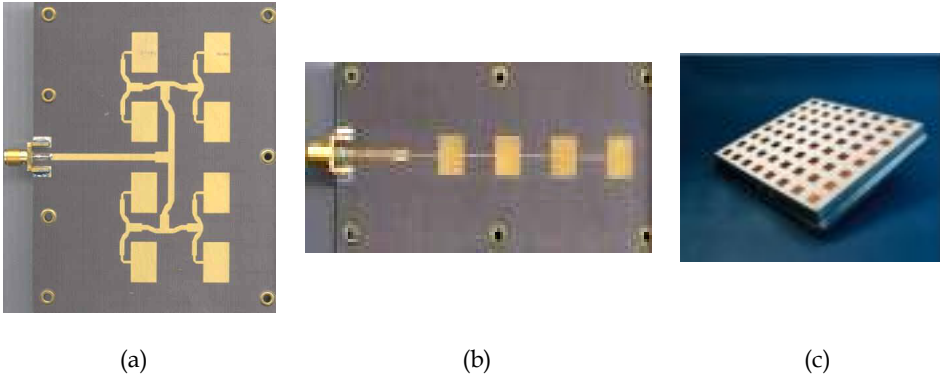
Fig. 4. Microstrip antenna array (a) phased array, (b) series-fed array, (c) Multi-Element Array (MEA)

4. Compact and Multiband Microstrip Antennas

4.1 Introduction to compact and multi band antennas

In wireless communication applications, the demand for low profile compact size planar antenna is increasing day by day. The light weight, small size handheld wireless devices design challenges never stops. Various types of low-profile elements have recently been developed and they are fairly efficient radiators that can be easily manufactured at low cost. The conventional microstrip patch is not a good candidate for the mobile wireless applications, due to certain disadvantages as narrow bandwidth. Therefore, more unusual approaches are investigated for multiband antenna with reduced size operation. Different other microstrip structures are successful candidates as microstrip Planar Inverted F-Antenna (PIFA) with different geometrical radiator shapes. The main goal is to design antennas for wireless communication applications where the space value of the antenna is quite limited while it reserves the characteristics of multiband, light weight, low cost, robustness, diversity, packaging capabilities and ability for MEMs integration for smart antenna systems. Several researches in literature concentrate on these antenna types and their developments. In this part the author will try to highlight the important results and several corresponding shapes.

Famous techniques for antenna size reduction include dielectric loading to reduce the electrical size, top hat loading, and use of shorting pins or plates are summarized in table 2 (Dalia Nashaat et al, 2003). Dielectric loading usually is accompanied by bandwidth reduction and cost increase, so it is not a likely approach. Thus, we will investigate antennas with top hat loading and shorting pins or plates, either separately or in combination.
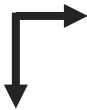
Performance Down sizing technique	Geometrical (size reduction with respect to conventional resonant size)	Circuital (impedance bandwidth enhancement ratio)	Radiation pattern (symmetry vs. low cross polarization)
Material: high dielectric, slow wave, etc.	70%	2%	high
Shorting planes: Walls, pins, etc.	50%	5%	low
Patch etching: slits, slots, notches, etc.	50%	7%	medium
Multilayered: fold, bend, stack, etc.	60%	10%	high

Table 2. Downsizing techniques and performance affecting on antennas

4.2 Compact/small antenna definition

The miniaturization of antenna structure is usually limited by physical laws of nature. Usually, the size reduction is on the expense of other antenna characteristics as bandwidth (BW), gain (g), etc. For small size antennas, there is always a tradeoff among antenna radiation quality factor (Q), BW and efficiency (η)(C. P. Huang, 1999; Dalia Nashaat et al, 2003).

The rule of thumb is:

$$\frac{BW\eta}{V} = \text{Constant} \tag{1}$$

Where BW is the antenna bandwidth, η is the antenna efficiency and V is the antenna volume.

$$BW\eta \cong (Ka)^3 \cong 1/Q \tag{2}$$

Where (a) is a radius of sphere that can include the small antenna. Form Equation 2, it indicates that the smaller the antenna, the higher the Q

So the microstrip antenna can be considered small if its largest dimension is less than or equal to (a) where:

$$a = 1/k = \lambda/2\pi \tag{3}$$

Where λ is the wavelength of the electromagnetic fields radiated by the antenna.. For example, for antenna operates at 5GHz, λ =60mm so a $\cong$ 10mm, which means that the antenna can be considered small if its dimensions are less than 10mm.

4.3-PIFA as compact multiband antenna

PIFA is well-known as terminal antenna design. These antennas offer reduced size over traditional microstrip antennas because the resonance frequency is at about quarter wave rather than at half wave in conventional ones due to the shorting pins/walls in its structure as shown in figure 5 (T. Taga, 1992).
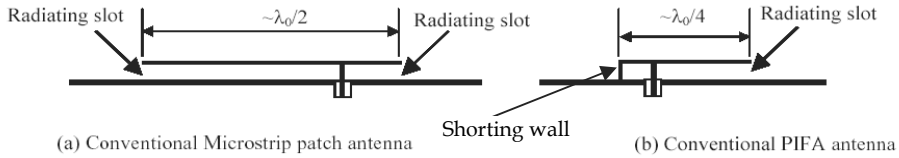
Fig 5. Comparison between conventional microstrip patch antenna and conventional PIFA antenna

The selection of PIFA is due to certain advantages as

The PIFA bandwidth is affected very much by varying the size of the ground plane, generally, reducing the ground plane can effectively broaden the bandwidth of the antenna system.

PIFA impedance matching can be obtained by the correct positioning of feeding and grounding pins. Thickness of the antenna and permittivity of the substrate material used also affect the impedance of the feeding point. To shrink the size of the PIFA, high constant dielectric substrate materials can be used. This weakens the performance of the antenna, because dielectric material gathers electromagnetic fields and therefore it doesn't radiate as good as the air insulated PIFA. Also part of the feed power goes into the dielectric losses of the substrate material. The height of the PIFA is a very critical dimension since it has a great effect on the antenna's radiation and also its impedance bandwidth (J. Elling et al, 1991; C. R. Rowell & R. D. Murch 1997). The basic rule is that the bigger the air gap between the radiator and ground plane is, the better the gain and the broader the impedance bandwidth will be. Table 3 summarizes the effect of different PIFA design parameters,(height, width, length, location of feed and shorting pin/wall and size of the ground plane) on its characteristics.

Parameters	Effect
Height	Control bandwidth
Width	Control impedance matching
Length	Increase inductance of the antenna and determine resonance frequency
Width of short strip	Affect on the anti-resonance and increase bandwidth
Feed position from short strip	Affect on resonance frequency and bandwidth

Table 3. The effect of PIFA parameters on its characteristics

4.4 PIFA structures for multiband and compact size applications:
4.4.1 Rectangular PIFA shape with U-shaped slots

A practical method to design a single feed multiband PIFA that covers both the cellular and non cellular bands is developed (Dalia Nashaat et al, 2005; Hala Elsadek, 2005; R. Chair et al, 1999). From the commercial point of view, there are now different frequency bands for portable cellular/non cellular devices as the conventional 0.9GHz GSM band for mobile phones and 1.8GHz DCS band for wireless cellular applications. Furthermore the Bluetooth wireless technology at 2.4 GHz is already applied in many portable devices and in most wireless communication systems as mobile phones, laptops, PDAS, car stereos, audio speakers, toys, etc (Bluetooth information web site). Moreover the band of WLAN at 5.2GHz is being applied in some applications. The compact and multiband functionality is not the only required demand in such antenna systems for wireless communication applications but, also other characteristics should be satisfied as small size, light weight, omni directional radiation pattern, reasonable gain and acceptable bandwidth.

Quad band PIFA with single coaxial probe feeding is investigated. Foam substrate is used for light weight, rigid structure and easy shielding purposes. Three U-shaped slots are added with certain dimensions and at appropriate positions for operation at the aforementioned four frequency bands. The size reduction is 30% from conventional quarter wavelength PIFA. Additional reduction by 15% is achieved by adding a capacitance load in the vertical direction. The impedance bandwidth is fairly acceptable. The antenna gain is satisfactory and the radiation pattern is quasi isotropic at the respective four bands of interest. The proposed concept of adding U-shaped slots is a distinct advantage of the design since the bands of operation are independent on each other except the small controllable mutual coupling between the slots. Figure 6 illustrates the suggested antenna design.

(a)
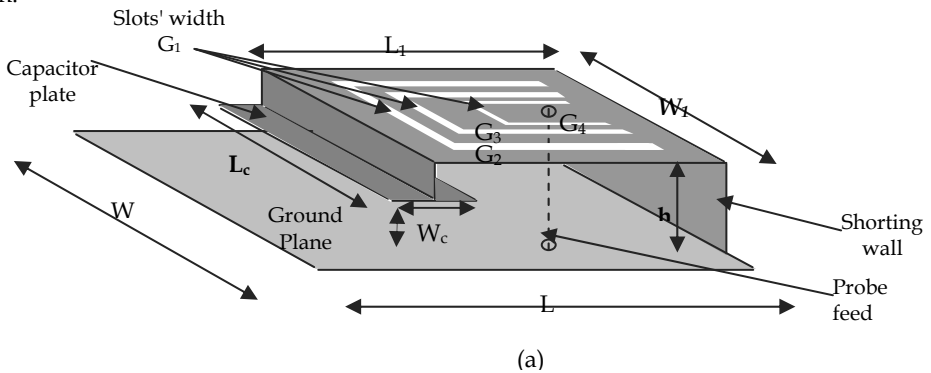
Fig. 6. Geometrical dimensions of the fabricated quad band antenna

The rule of thumb in antenna design is:

$$f_i = \frac{c}{4(L_i + W_i)} \tag{4}$$

The length L_i and width W_i are replaced by L_1 and W_1 =(61mm,40mm) of the PIFA rectangular radiating surface to determine the first resonance frequency f_1 (0.9GHz). While, (L_i, W_i) are replaced by the dimensions of the largest U-slot (L_2, W_2)=(23mm,30mm) to

generate the second resonance frequency f_2 (1.8GHz). They are also replaced by the length $(L_3, W_3) = (18mm, 20mm)$ of the middle U-slot to get the third resonance frequency f_3 (2.45GHz). Finally, (L_i, W_i) are replaced by $(L_4, W_4) = (9.5mm, 8mm)$ of the smallest U-slot to have the fourth resonance frequency at f_4 (5.2GHz). This multi-band antenna has approximately the same size as a single-band PIFA operating at the lowest frequency band. The radiating element is grounded with a shorting wall. It is found that the widest bandwidth is achieved when the width of this wall is equal to the width of the PIFA radiating plate. The antenna is fed using coaxial cable at the appropriate matching point for the four bands of operation. The antenna impedance can be matched to 50Ω by controlling the distance between the feed point and the shorting wall. The PIFA antenna is fabricated on a foam substrate with dielectric constant $\varepsilon_r = 1.07$ in order to have rigid structure that can be easily shielded. Adding U-slots on the PIFA radiating surface, reduces its size by about 30% from the conventional PIFA shape. For further reduction in size, a capacitor plate load is added between the radiating surface and the ground plane. This increases the reduction in size to be about 45%. The results of the structure simulations as well as experimental measurements are illustrated in following three figures.
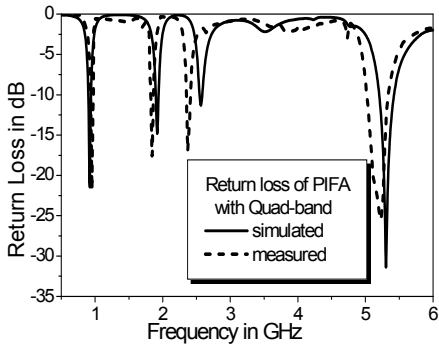
Fig. 7. Comparison between measured and simulated reflection coefficients of quad band PIFA with three U-shaped slots at operating frequencies of 0.95, 1.8, 2.45 and 5.2GHz, respectively.
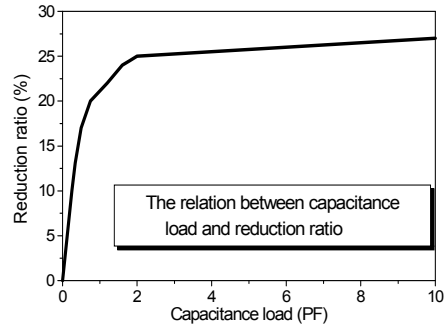
Fig. 8. The relation between capacitor load in PF and antenna percentage reduction ratio compared to conventional PIFA.
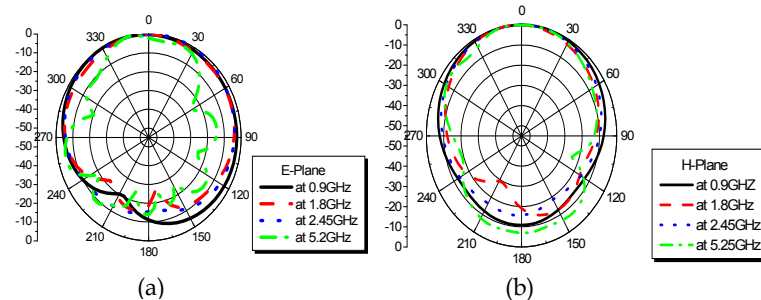
Fig. 9. The simulated radiation pattern of quad-band PIFA with 10PF shorting capacitor plate at four different resonating frequencies, a) at parallel E-plane at phi=0 and b) at perpendicular H-plane at phi=90.

4.4.2 Compact PIFA size with E-shaped radiator

Ultra compact PIFA with dual band resonant frequencies are investigated (Hala Elsadek, 2006). The antenna is designed and fabricated on both foam and FR4 cheap substrates with dielectric constants ε_r = 1.07 and 4.7, respectively. Over 95% reduction in the antenna size is achieved from conventional $\lambda_0/4$ rectangular PIFA resonating at same frequencies. This is done by implementing two oppositely shorting capacitive straps under the radiating surface. Dual band operation is achieved by inserting two parallel slots on the edges of the PIFA radiating surface forming an E-shape. In this case, the center wing resonates at the higher frequency while the two side wings resonate at the lower frequency. The antenna resonance frequencies on FR4 substrate are 1.07GHz and 2.77 GHz with areas' reduction ratios of 97% and 81% for the lower and upper resonance frequencies, respectively. The antenna size on FR4 substrate is 13 x 11 x 8mm³. The antenna directivity is 3.73 with radiation efficiency 97%. The radiation pattern has acceptable shape with low cross polarization in both resonances and at both E-plane and H-plane directions. It is worth to mention that, with frequency scaling, the same antenna structure can resonate at 2.4GHz and 5.2GHz with dimensions 8mmx8mmx8mm Figure 10 shows the antenna geometry, while figure 11 illustrates a comparison between simulated and measured results with capacitive load reduction effect. There are different approaches for multiband compact antenna design; however, we concentrated on PIFA with shorting plates and capacitive loads with different radiator shapes. Since these shapes give excellent results for antenna candidates in mobile communications.
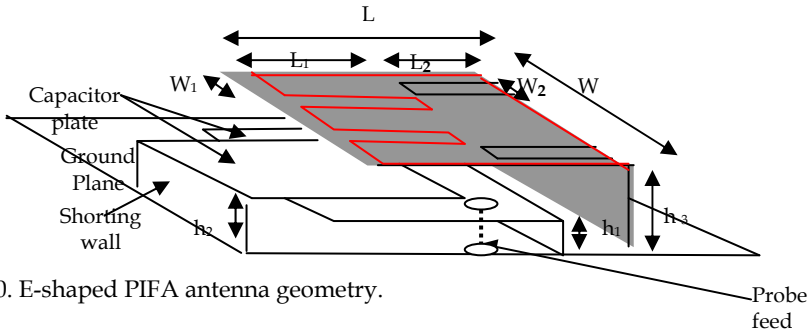
Fig. 10. E-shaped PIFA antenna geometry.

(a)   (b)
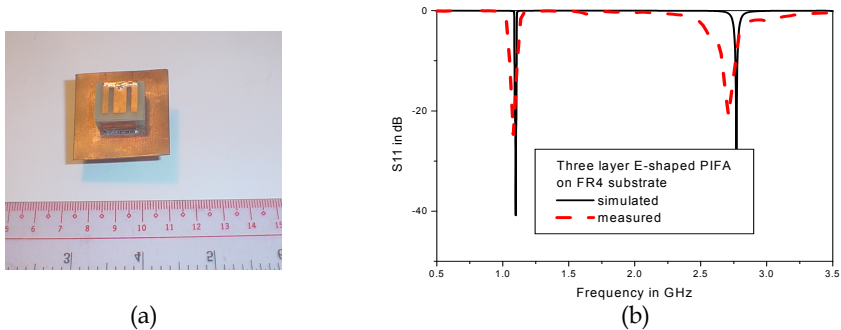
Fig. 11. Photo of fabricated E-shaped PIFA on commercial FR4 substrate and (b) comparison between simulated and measured antenna reflection coefficients.

5. Broad band and UWB Antennas

5.1- Introduction to broad band and UWB antennas

In last sections, we illustrate the challenge of small and multiband antenna that can fit in several wireless communication systems at same time. In all previous designs, acceptable antenna bandwidth was achieved. However, several other applications of wireless communications require broadband and even ultawideband antenna rather than directional one. Broad band antennas are desired for the increasing demand of communication bandwidth that accommodates high data rate application like video-on-demand. Moreover UWB technology attracts a lot of attention from the researchers in recent years because of the various advantages it offers. UWB technology depends on transmitting pulses of width in order of nano seconds instead of modulating sinusoidal signal and, hence broadening its spectrum and tuning its power density beyond noise level (FCC, 2002). This method in transmission exhibit many advantages as immunity to jamming and ability to combat fading due to multipath effects. Also it has penetration capability as its spectrum include low frequency components. Because of these advantages UWB technology has enormous applications in wireless communications. One of the major application is the wireless sensor network (WSN) which is useful in medical, tracking and localization applications (remote sensing) (Ian Opperman at el., 2004; K.P. Ray, 2008). As UWB provide security and low power consumption that increase the battery life of the portable terminals. On the other hand, broad band communication systems as well as UWB technology faces a lot of challenges as the radiation pattern stability and polarization purity along the whole band of operation.

5.2-Different types of broad band antennas

Many designs have been investigated in literature for broadening the bandwidth of antennas. This can be achieved by using different probe feeding shapes as L-shape, adding parasitic elements to the radiator, folding the ground plane, etc. (Fan Yang, 2001; Yasshar Zehforoosh, 2006). Taking in consideration for the stability of the beam pattern and polarization purity along the bandwidth, the design quality is judged. Among the basic ideas for broadening the band are inserting slots of different shapes (U,H,V) on the radiating patch antenna to introduce longer current paths and hence add other staggered resonating modes. The rule of thumb in adding another resonance to the antenna structure is the same as that discussed in previous section for multiband antenna designs however, in case the resonating modes are far from each other, the structure will act as multiband antenna. But if the design is changed to let these resonances near from each other, they will complement each other forming staggered resonating behavior and broadband antenna structure. Also adding parasitic or stacked patch has been proposed in (Mohamed A. Alsharkawy at el., 2004). Another types as aperture stacked and multi resonator stacked patches in (Ki-Hakkim at el, 2006; Jeen Sheen Row, 2005) .In these types multi patch antenna are printed on different layer forming multi resonators and hence broaden the antenna band. These types are bulky and not adequate enough to be integrated with the modern wireless devices in spite there are successful attempts for this. In addition they don't exhibit enough bandwidth to cover all wireless communication band nowadays (3.1-10.6GHz). Recently UWB slot antenna in (Girish kumar, 2003; Yashar Zehforoosh at el, 2006) and printed monopole antenna in (Soek H. Choi at el., 2004) are proposed. They attract a lot of interests due to their

low profile, ease of integration and very wide bandwidth. Next section will focus on the UWB printed monopole antenna.

5.3- UWB antenna Design

Some considerations should be taken for UWB antenna design such (Hung-Jui Lam, 2005):
1-It should have bandwidth ranging from 3.1GHz to10.6GHz in which reasonable efficiency is satisfactory.
2-In this ultra-wide bandwidth, an extremely low emission power level should be ensured. (In 2002, the Federal Communication Commission (FCC) has specified the emission limits of −41.3 dBm/MHz).
3-The antenna propagates short-pulse signal with minimum distortion over the frequency range.

5.4 UWB Printed Monopole Antenna

Printed monopole antenna structure is shown in Figure 12 and it could be explained as an evolution of the conventional microstrip antenna with ground plan eliminated (K.P. Ray, 2008). From the analysis of the microstrip antenna, (Hirasawa and K. Fujimoto, 1982; C.A. Balanis, 1997) it is known that the substrate thickness (h) is directly proportional to the BW and as (h) is extended to infinity by eliminating the ground plan the BW become very wide. Also, the resonant frequency is function of the patch length, width and height. So when patch printed on very thick substrate it excites higher order modes each enables broad
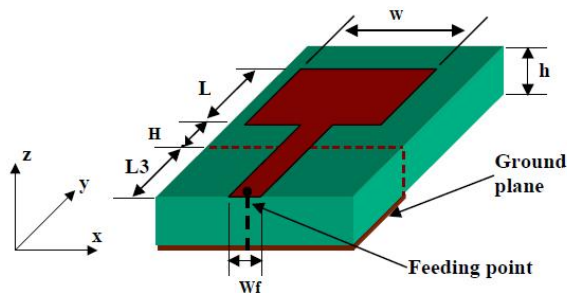
Fig. 12. Geometry of the rectangular printed monopole antenna

bandwidth case. If these higher order modes are close to each other the overall bandwidth is ultrawideband. Another explanation for the printed monopole that it could be seen as conventional monopole but with the cylindrical metallic rod flatted to be plane of any different shapes (K.P. Ray, 2008) (rectangular, circular, elliptic)as it is known that impedance bandwidth increase by increasing the diameter of the metallic rod. The printed plane that alternate the metallic rod is considered of diameter extended to infinity exciting higher order modes of large bandwidth. Upon optimizing the dimensions of the antenna, these higher order modes could be close to each others to yield very broad bandwidth as will be elaborated in next sections.

5.4.1 Analysis

As mentioned in previous section, printed monopole antenna is analog to the wire quarter wave monopole antenna. This could be used to analytically design the antenna for the lower edge frequency by equating its area (in this case rectangular monopole) to an equivalent cylindrical monopole antenna of same height L and equivalent radius r as following:

$$2\pi r L = WL \tag{5}$$

The input impedance of thin $\lambda/4$ monopole is half the input impedance of thin $\lambda/2$ dipole and equal is slightly less than quarter wavelength and given by(15, 38)

$$
\begin{aligned}
&L = 0.24 \times \lambda \times K \\
&where \\
&K = (L \ / \ r) \ / \ (1 + L \ / \ r) = L \ / \ (L + r) \\
&\lambda = \frac{(L + r)}{0.24} \\
&therefore \\
&f_l = \frac{c}{\lambda} = \frac{(30x0.24)}{L + r} = 72 \ / \ (L + r)...GHz
\end{aligned} \tag{6}
$$

Previous equation doesn't account for the distance between the radiator and the ground plane (h)

$$f_l = 72 \ / \ (L + r + h)...GHz \tag{7}$$

where all dimensions are in millimeters. This analysis is valid for free space but in our case where antenna is printed on a dielectric substrate which decrease the effectiveness of the wavelength (λ_g). Modification on the lower edge frequency is required and can be given by

$$f_l = 72 \ / \ (L + r + h) \times k.GHz \tag{8}$$

It is worthwhile to mention although previous analysis was on rectangular shape printed monopole, it is valid on other various shapes of radiators but only L and r will differ according to the geometry of the shape. (K. P. Ray, 2008).

After inspecting the lower edge frequency we need to control the bandwidth of the antenna. Actually the L, r and h affects both lower edge frequency as well as the bandwidth too so optimization is needed to give the required bandwidth as well as the lower frequency. Another important thing that affects severely the bandwidth is the bottom shape of the radiator in contact with the 50Ω feeder. As long as we avoid abrupt change in the dimensions of the transition from the feeder to the radiator as long as we obtain broader bandwidth. That's why circular radiator inherent wider band than rectangular one. Abrupt transition form feeder to radiator is overcome by using stepped or tapered feeders (S. I. Latif at el., 2005; A.P. Zhao and J. Rahola, 2005). Finally using CPW (coplanar waveguide feed) instead of microstrip feed enhances the bandwidth. As printed monopole antenna resonating around quarter wave length so they have similar radiation pattern as normal

monopole. It is omni in the H-plane and eight shaped in the E-plane. Following are examples about broad band and UWA antenna designs.

5.5 Examples on braodband and UWB microstrip antenna designs
5.5.1 Broad band antenna

The geometry of the proposed antennas is as shown in figure 13. The antenna consists of V-shaped patch with V- unequal arms with dimensions (L_1, W_1) and (L_2, W_2). The isosceles triangular antenna is with dimensions (L_T, W_T). The shorting wall width is equal to W_T for maximum size reduction (Hala Elsadek and Dalia Nashaat, 2008). The ground plane is with rectangular shape of dimensions (L_g, W_g). The two parts of the structure, V-shaped patch and triangular PIFA, are coupled through a V-shaped slot with unequal arms with slots' lengths and widths are (L_{s1}, W_{s1}) and (L_{s2}, W_{s2}). The two arms of the V-shaped patch excite TM_{01} mode. The length of the two arms of the V-shaped patch is different in order to excite two different staggered resonant modes. The unequal spacing/widths between the coaxially fed triangular shorted patch and the V-shaped patch are for different values of coupling thus, excite two more different modes. To add two more resonating modes, equal arms V-shaped slot can be loaded on the triangular patch radiation surface. The substrate is foam with dielectric constant ε_r =1.07 and substrate height h=6mm. The antenna geometry is illustrated in figure 13. When the ground plane size is reduced to certain proper value, the antenna behavior changes to be wide bandwidth antenna rather than multiband antenna.

The resonating frequencies can be approximately determined from following equation (Yujiang Wu and Zaiping Nie, 2007).

$$f_i = \frac{c}{4L_i} \qquad (9)$$

Where: f_i is resonant frequency at band i, C is the velocity of light $= 3 \times 10^8$ m/s and L_i is the half length of the radiating surface or the length of the slot at the corresponding operating band i. The Triangular PIFA part is excited by coaxial probe feed. The probe is positioned in the centerline of the shorted patch at distance d_f from shorting wall. The d_f value controls the antenna characteristics. For multiband operation, the resonating frequencies are at 2.88GHz, 3.64GHz, 3.95GHz, 4.38GHz, 4.81GHz and 5.6GHz, the distance d_f is 16.75mm while for broadband operation, the distance d_f increased to be 18.5mm.

Figure 14 illustrates comparison between the simulated and measured results for the multiband structure. The radiation pattern of the antenna is approximately omni directional in both E-plane and H-plane with back to front ratio of less than 5dB and 3dB beamwidth of about 60°.

Moving coaxial feeding towards open end of triangular PIFA antenna at d_f = 18.5mm, the resonant frequencies of the antenna become staggered close to each other so achieving wideband operation. The bandwidth is 3% at the fundamental mode 2.95 GHz, hence the fundamental resonating frequency will approximately not affected by changing the feed

position. The higher resonance bandwidth is 27% at 4.721GHz. Figure 15 presents the comparison between the simulated and measured results of the wideband antenna structure.

Folding the shorting wall of the triangular PIFA as in figure 13, converts the antenna to UWB with bandwidth of 53% at same resonating frequency 4.65GHz. The antenna gain is 10.5 dBi
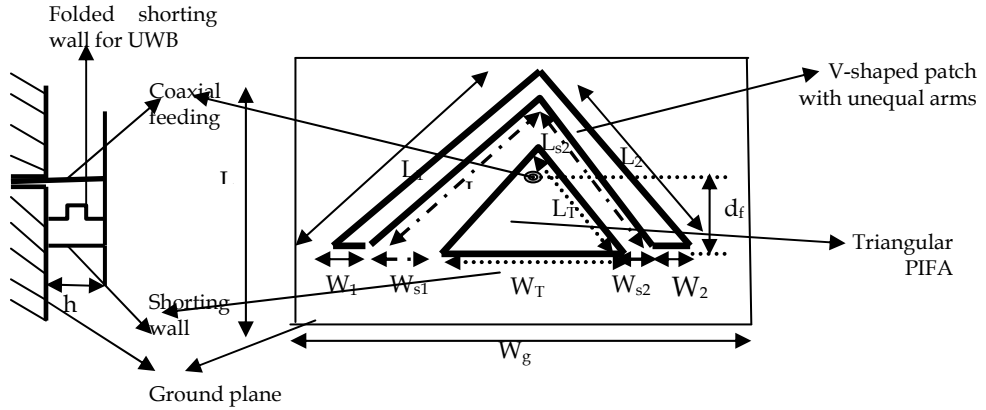
Fig. 13. Configuration of the proposed antenna of V-shaped patch with unequal arms coupled to isosceles triangular PIFA through V-shaped slot of unequal arms
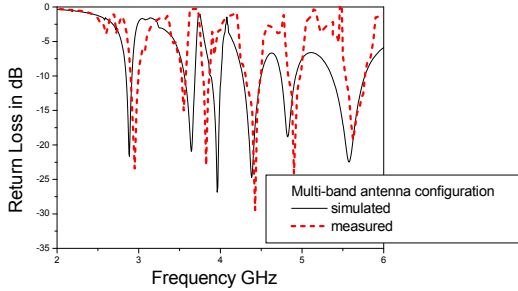
Fig. 14. Comparison between simulated and measured results of the multi-band antenna
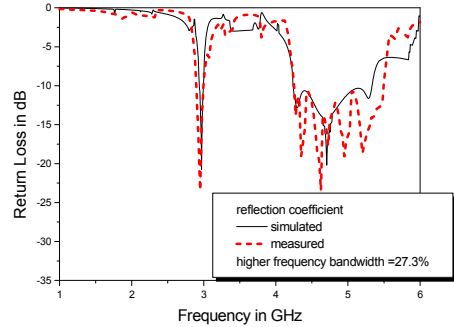
Fig. 15. Comparison between simulated and measured results of broad band antenna

5.5.2 UWB antenna

Consider we have substrate material of ε_r =3.38 and h=0.813mm and we need to design printed rectangular monopole shown in figure 12 so we need to know the values L,W,H for obtaining lower edge resonance frequency at 5Ghz and obtain BW as Wide as possible. From above equations in subsection 5.4.1, to satisfy 5GHz a lot of solutions could be obtained for L, W, h but not all of them will give the maximum BW, so optimization is

needed for obtaining the optimum dimensions. Parametric analysis for the effects of these three dimensions on bandwidth is shown in figures 16-18 (Hakim Aissat at el, 2006; Min Hau Ho at el, 2005). Starting with L=W=0.25λ_0/$\sqrt{\varepsilon_r}$ =8mm and h=2mm. From the three figures below, the optimum dimensions are W=12,L=11.5 and H=0.75.
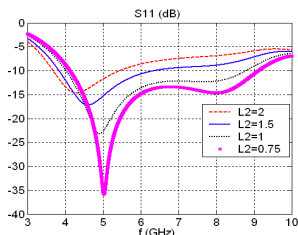
Fig. 16. The effect of changing *W* on the return Loss at *L*=5 mm and *h*=2 mm.
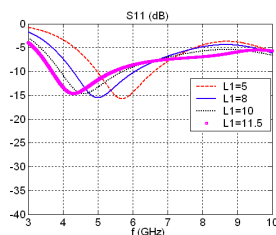
Fig. 17. The effect of changing *L* on the return loss at *W*=12 mm and *h*=2 mm.
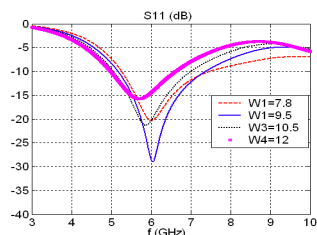
Fig. 18. The effect of changing *h* on the return loss at *W*= 12 mm and *L*= 11.5 mm.

6. Reconfigurable microstrip antenna

6.1 Introduction to reconfigurable antenna system

Due to the increasing demand of multipurpose antennas in the modern wireless communication devices and radar systems, reconfigurable antennas have attracted a lot of researcher's attention. One type of these antennas capable for operation at mutli bands and hence could intercept various communication systems (KPCS/WiMAX/GSM/WCDMA) with lower co-site interference. Other types exhibit diversity in transmission or reception to combat fading effects and enhance signal quality. Reconfigurable antennas are similar to the conventional antennas but one or more of its specification or characteristics could be adjusted or tuned using RF switches/MEMs or variable capacitors/inductors. They have four types: 1-Frequency reconfigurable, 2-poalrization diversity, 3-radiation pattern steering, 4-combination of the three previous types. Advantages of reconfigurable antennas are integration with wireless and radar devices instead of multiple antenna systems, compactness, cost reduction, etc. Frequency reconfigurable antenna could decrease interference and make efficient use of the electromagnetic spectrum. Polarization diversity and radiation pattern steering antennas could lead to increase in the communication system capacity and fading immunity. Moreover they open the way of emerging some modern communication systems like MIMO and cognitive radio. Also from future potential for the introduction of smartness and intelligence to the handheld terminals. Switching and/or tuning takes place with the aid of PIN diodes or MEMs switches or varactors adopted with the antenna structure. Pin diodes are reliable and experience high switching speed but introduce non linearity and need complex bias circuitry to be integrated with the antenna. On the other hand MEMs have lower insertion loss, easier in integration (no need for biasing circuitry), less static power consumption and have higher linearity, but it needs high static bias voltage. According to the various advantages of reconfigurable antennas they are currently part of many modern wireless communication systems such as (DCS/GSM/WCDMA/Bluetooth/WLAN), hand held GPS and other navigation systems,

MIMO Systems and steerable arrays. In the following different examples and kinds of reconfigurable antennas will be presented.

6.2 Polarization Diversity

The work proposed by Hakim Aïssat in (Hakim Aissat at el, 2006) provide circular antenna with switchable polarization as shown in Figure 19 in which 2 diodes in the ground 45^0 slot are ON and the others in the 0^0 slots are OFF make the antenna linear polarized on contrary, when the other 2 diodes in the 45^0 are ON and in the 0^0 are OFF make the antenna circular polarized. Thin slits (130um) were made in the ground plane to avoid DC short on the switches. A rule of thumb for the switches biasing circuitry design that the RF shouldn't go to the DC and the DC shouldn't affect the RF. So for DC blockage from RF, large capacitors are built over the slits by stacking copper strips and adhesive tapes (upper layer on Figure 19). The slits are first covered by an isolating adhesive layer, which insures a dc isolation maintaining RF continuity. The adhesive layer is then topped with four copper tapes to shield the slits at RF frequencies. This antenna enables diversity in TX/RX and hence could enhance signal quality or increase system capacity by polarization multiplex.

For other shapes of microstrip antenna in (Yujiang Wu & Zaiping Nie, 2007) proposed square patch with switchable polarization RHCP/LHCP using 4 pin diodes. The antenna is shown in Figure 20.Circular polarization is synthesized by truncating two opposite corners of the patch. And both LHCP/RHCP are generated by double feeding the patch from two orthogonal sides. Switching ON/OFF diodes 1&2 shown in the Figure 20 in opposite manner achieve RHCP and LHCP, respectively. Also linear polarization could be obtained by attaching triangular small strips connected to the truncated corners and connecting them to the patch via pin diodes3&4 as shown in Figure 20. When these diodes are ON linear polarization is exhibited.
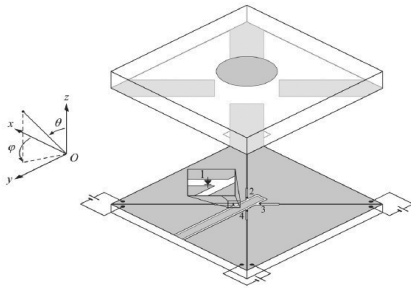
Fig. 19. Circularly polarized reconfigurable antenna
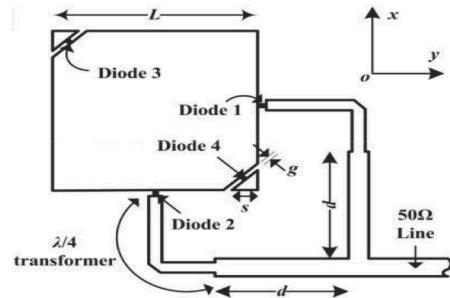
Fig. 20. Configuration of the corner-truncated square microstrip antenna with switchable polarization

6.3 Radiation pattern Steering

Adaptive beam spiral antenna found in the work done in (Greg H. Huff et al, 2004). The geometry of the antenna is shown in Figure 21 where positioning of open circuit in the spiral arm change current distribution leading to steering the beam direction. Two switches are used to open/close the open circuit in the spiral arm and hence the pattern direction is two bit controllable.

In (Yong Zhang et al, 2005), a fractal Hilbert microstrip antenna with reconfigurable radiation patterns using 8 switches is proposed. The antenna is shown in Figure 22. By turning switches on and off interesting results can be obtained. For example at switch pairs (a3, a4) & (a7, a8) are OFF and the others, (a1, a2) & (a3, a4) are ON and then alternates between the status of the maximum radiation patterns in E-plan steer from +/-30^0 among two states. The radiation patterns in the H-planes of the two states are almost the same.
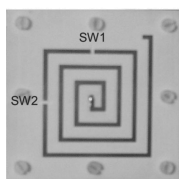
Fig. 21. reconfigurable rectangular spiral antenna (SPRL)
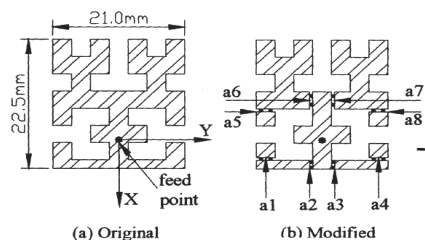
Fig. 22. Geometry of the original and modified reconfigurable fractal Hilbert microstrip antennas

6.4 Frequency reconfigurable antennas

(Ahmed Khirde & Hala Elsadek,2009) proposed simple design for a low cost band notch UWB printed monopole antenna with reconfigurable capability in a way you are able to create and cancel band notch in the UWB spectral mask covering the in-band IEEE802.11a/h co-existing systems. The antenna dimensions are shown in Figure 23(a) and Figure 23(b). A patch is added in the back plane as a parasitic half wave resonator coupled electrically to the rectangular monopole. Slot is cut into the parasitic patch as shown in Figure 23(b).The two patch parts are connected by means of two RF switches S1 and S2 which are modeled as metal pad with dimension 0.3x0.9mm.Although this model is ideal, it gives a very good approximation for the real commercial pin diode switch HPND-4005 manufactured by HP. The role of the switches here is to reconfigure the antenna between the ON/OFF states. The ON state where the antenna exhibit a band notch covering the bandwidth of the WLAN for IEEE 802.11a/h which is 5.15-5.825 GHz. The OFF state the band notch is removed and the antenna bandwidth returned flat. The simulation and experimental results comparison for the ON state and OFF state are presented in figures 24 and 25, respectively. The antenna gain over the whole band is presented in figure 26.
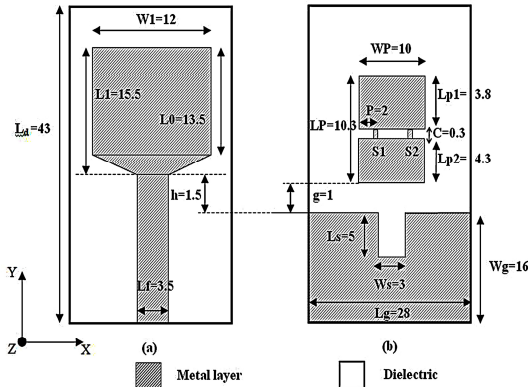
Fig. 23. Geometry of band notch monopole antenna
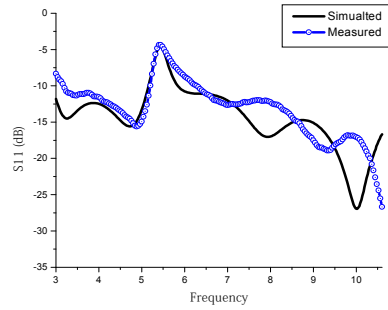
Fig. 24. Comparison between simulated and measured return loss of the ON state
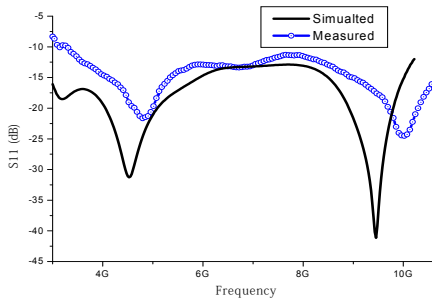

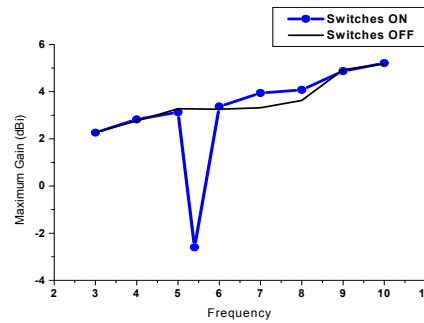Fig. 25. Comparison between simulated and measured return loss of the OFF state

Fig. 26. Antenna gain in the on and off states

As mentioned above reconfigurability could be achieved using RF Micro Electro-Mechanical (MEMs) switches or actuators. An example for a frequency tunable antenna suing MEMs micromachining is proposed in (R. Al-Dahlehet al,2004). It is simple patch printed on Silicon using VLSI micorelectronics technology and Air gap is beneath the patch. .MEMs actuator is to change the thickness of the air gap beneath the patch, hence changing the effective substrate dielectric so the resonance frequency is changed. The antenna structure is shown in Figure 27. This kind of antennas is very important for antenna on chip and modern Soc technology that are vital for compact handheld devices.
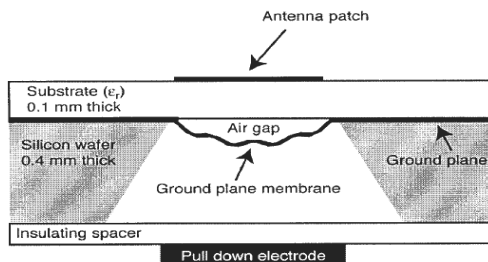
Fig. 27. Schematic of the frequency tunable microstrip patch antenna showing the MEMs membrane in the ground plane below the antenna

7. Smart Microstrip Antennas

7.1 Intorduction to smart antenna system

A smart antenna system consists of either single antenna element or combines multiple antenna elements with a signal processing capability to optimize the radiation and/or reception pattern automatically in response to the required signal environment. Different technologies are combined and defined today as smart antenna system. These ranges from simple diversity antennas to fully adaptive antenna array systems. In truth, antennas are not smart by itself—antenna systems are smart. In other words, such a system can automatically change the directional of its radiation patterns or any other characteristic like resonating frequency, polarization direction, antenna gain, antenna bandwidth, etc. in response to its surrounding signal environment. This can dramatically improve the performance (such as capacity and coverage range) of the wireless system.

7.2 Smart antenna systems classifications

Sectorization schemes, which attempt to reduce interference and increase capacity, are the most commonly spatial technique that have been used in current mobile communication systems for years. Cells are broken into three or six sectors with dedicated antennas and RF paths. Increasing the amount of sectorization reduces the interference seen by the desired signal. One drawback of the sectorization techniques is that the efficiency decreases as the number of sectors increases due to antennas' patterns overlap. Any reduction in the interference level translates into system capacity improvements. Smart antennas could be divided into two major types, fixed multiple beams and adaptive array systems. Both systems attempt to increase gain in the direction of the user. This could be achieved by directing the main lobe, with increased gain, in the direction of the user, and nulls in the directions of the interference (Ahmed Elzooghpy The international engineering consortium; 2005).

The following are distinctions between the two major categories of smart antennas regarding to the choices of transmit strategy:

Fixed multiple switched beam: A finite number of fixed, predefined patterns or combining strategies (sectors) are transmitted

Adaptive array: an infinite number of patterns (scenario-based) are adjusted in real time.

7.2.1 Switched beam smart antenna system

Switched beam antenna systems form multiple fixed beams with heightened sensitivity in particular directions. These antenna systems detect signal strength, choose from one of several predetermined, fixed beams, and switch from one beam to another as the user moves throughout the sector. In terms of radiation patterns, the switched beam approach subdivides macrosectors into several microsectors as a means of improving range and capacity. Each microsector contains a predetermined fixed beam pattern with the greatest sensitivity located in the center of the beam and less sensitivity elsewhere.

7.2.2 Adaptive smart antenna system

Adaptive antenna technology represents the most advanced smart antenna approach to date. Using a variety of new signal-processing algorithms, the adaptive system takes advantage of its ability to effectively locate and track various types of signals to dynamically minimize interference and maximize intended signal reception. Adaptive arrays utilize sophisticated signal-processing algorithms to continuously distinguish between desired signals, multipath, and interfering signals as well as calculate their directions of arrival. This approach continuously updates its transmit/receive strategy based on the changes in both the desired and interfering signal locations (Ahmed Elzooghpy, the international engineering consortium, 2005). Figure 28 illustrates comparison between the two smart antenna systems coverage.
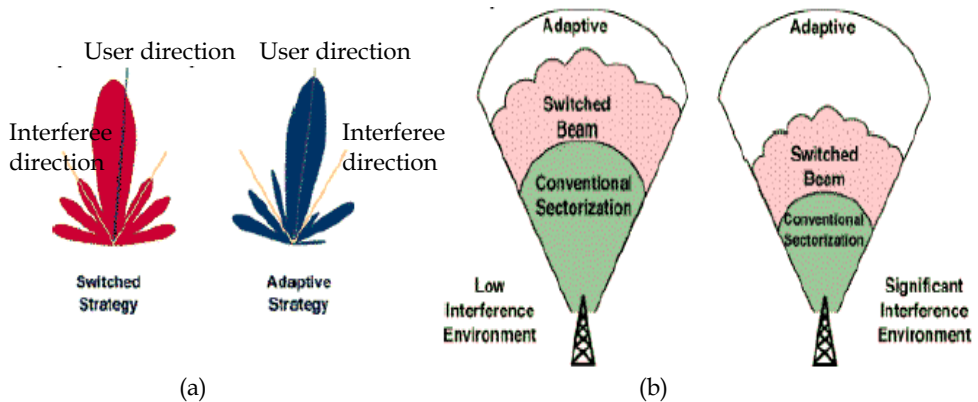
Fig. 28 (a) Beam forming lobes and nulls in switched and adaptive array systems, green lines are the required user direction and yellow lines are for co-channel interference and (b) coverage patterns for switched beams and adaptive array antennas

7.3 Advantages and disadvantages of smart antenna system
7.3.1 Advantages

The dual purpose of a smart antenna system is to augment the signal quality of the radio-based system through more focused transmission of radio signals while enhancing capacity through frequency reuse. The main advantages of the smart antenna system and their reflected effect on system performance are listed in table 4 below (Michael Chryssomallis, 200; Rappaport, T. S., 1998; Tsoulos G. V., 2001).

Feature	Benefit
Signal gain: Inputs from multiple antennas are combined to optimize available power required to establish required level of coverage.	**Better range-coverage**: Focusing the energy increases the base station coverage range. Lower power requirements also enable a greater battery life and smaller/lighter handset size.
Interference rejection: Antenna pattern can be generated toward co-channel interference sources, thus improving the signal-to-interference ratio of the received signals.	**Increase capacity**: Precise control of signal nulls and mitigation of interference allows for improving capacity.
Spatial diversity: Composite information from the array is used to minimize fading and other undesirable effects of multipath propagation.	**Higher bit rates transfer**: multipath rejection reduces the effective delay spread of the channel which allows for higher bit rates to be supported
Power efficiency: It combines the inputs from multiple elements to optimize available processing gain in the downlink (toward the user)	**Reduce expense**: Lower amplifier costs, reduce power consumption, and increase reliability.

Table 4. Benefits of smart antenna system

7.3.2 Disadvantages

One of the major existing disadvantages of smart antennas is in their complex hardware design and implementation. Multiple RF chains can increase the cost and make the transceiver bulkier. Most of the baseband processing requires coherent signals. This means that all the mixer LOs and ADC clocks need to be derived from same sources. This can present significant design challenges. The phase characteristics of RF components can change over time. These changes are relatively static and hence need calibration procedures to account for phase differences.

7.4 Applications of smart antenna systems

Smart antenna technology can significantly improve wireless system performance and economics for a range of potential users. It enables operators of PCS, cellular, and wireless local loop (WLL) networks to realize significant increases in signal quality, capacity, and coverage.

Adaptive antennas have been used in areas such as radars, satellite communications, remote sensing, and direction finding, to name a few. For instance, radar and secure communications systems take advantage of the ability of these antennas to adapt to the operating environment to combat jamming. Satellite communication systems have used multiple beam and spot beam antennas to tailor their coverage to specific geographic locations. (Kawala P. and U. H. Sheikh, 1993).

7.5 Future perspectives for smart antennas systems

According to recent studies, smart antenna technology is now deployed in one of every 10 base stations in the world, and the deployment of smart antenna systems will grow by 60 percent in the next four years. It was shown in the same study that smart antenna

technology has been successfully implemented for as little as 30 percent more cost than similar base stations without this technology.

Smart antennas are already part of current releases of 3G standards and more sophisticated approaches are considered for future releases. Furthermore, there is currently increasing interest in the incorporation of smart antenna techniques for IEEE wireless LAN/MAN (802.11n and 802.162). However, implementation costs can vary considerably, and cost-effective implementation is still the major challenge in the field. At the base station of particular importance is the development of improved antenna structures (possibly employing MEMS, technology, e.g., micro-switches, or improved cabling structures, and efficient low-cost radio frequency/digital signal processing architectures. At the terminal, the application of smart antenna techniques can have a significant impact, in terms of system performance, cost and terminal physical size (T.L.Roach et al, 2007). The financial impact of the deployment of smart antenna technologies in future wireless systems was studied in (Angeliki Alexiou & Martin Haardt, 2006) for CDMA2000 and UMTS. The results showed that smart antenna techniques are key to securing the financial viability of operators' business, while at the same time allowing for unit price elasticity and positive net present value. They are crucial for operators that want to create demand for high data usage and/or gain high market share. Based on this type of analysis, technology roadmaps along with their associated risks can be concluded that enable appropriate technology intercept points will be determined, resulting in the development of technologies appropriate for each application area.

8. Acknowledgment

9. References

Ahmed Elzooghpy, "Smart antenna engineering", Artech House, Inc., Mobile communication series, 2005.

Ahmed Khidre, Hala Elsadek and Hani Fikry," Reconfigurable UWB printed antenna with band rejection covering, IEEE802.11a/h', IEEE Int. Symp. on antennas and propagation, South Carolina, June. 2009

Angeliki Alexiou and Martin Haardt," Smart antenna technologies for future wireless systems: trends and challenges", IEEE communication magazine, Sep. 2006, pp: 90-97.

A. P. Zhao and J. Rahola, "Quarter-wavelength wideband slot antenna for 3–5 GHz mobile applications," IEEE Antennas Wireless Propag. Lett., vol. 4, pp. 421–424, 2005.

Bluetooth information web site," www.anycom.com, Bluetooth@anycom.com".

C. A. Balanis, Antennas theory: Analysis and Design, second edition, John Wiley & Sons, USA, 1997, ch.2, 6, 7, 12, 14.

C. P. Huang, "Analysis and design of printed antennas for wireless communications using the finite difference time domain technique," Ph.D. Dissertation, Electrical Engineering Department, University of Mississippi, December 1999.

C. R. Rowell and R. D. Murch, "A capacitively loaded PIFA for compact mobile telephone handsets," IEEE Trans. Antennas Propagat., vol. 45, no. 5, pp. 837-842, May 1997.

Dalia Nashaat, hala Elsadek and hani Ghali, " Single Feed compact quad band PIFA antenna for wireless communication Applications", IEEE Transactions on Antenna and Propagation, Vol. 53, No. 8, Aug 2005, PP: 3631: 2635.

Dalia Nashhat, Hala Elsadek and Hani Ghali "A Wideband Compact Shorted Rectangular Microstrip Patch Antenna with U-Shaped Slot," will be published in the Proceedings of IEEE Antenna and Propagation Symposium, Columbus, Ohio, June 2003.

D. M. Pozar and D.H. Schaubert, "Microstrip Antennas The Analysis and Design of Microstrip Antennas and Arrays," IEEE Press, New York, 1995.

Fan Yang, Xue-Xia Zhang, Yahya Rahmat-Samii, "Wide-Band E-Shaped Patch Antennas for Wireless Communications", IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, VOL. 49, NO. 7, JULY 2001.

FCC, First Report and Order on Ultra-Wideband Technology, 2002.

Girish Kumar, K. P. Ray, Broad Band Microstrip Antennas, Boston. London: Artech House 2003

Greg H. Huff, Judy Feng, Shenghui Zhang, Garvin Cung, and Jennifer T. Bernhard, "Directional Reconfigurable Antennas on Laptop Computers: Simulation, Measurement and Evaluation of Candidate Integration Positions", IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, VOL. 52, NO. 12, DECEMBER 2004.

Hakim Aïssat, Laurent Cirio, Marjorie Grzeskowiak, Jean-Marc Laheurte, and Odile Picon, "Reconfigurable Circularly Polarized Antenna for Short-Range Communication Systems", IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, VOL. 54, NO. 6, JUNE 2006

Hala Elsadek," Clip Antenna for Wireless Blutooth Applications ", IEEE Antenna and Propagation Magazine" Vol. 47, No. 3, June 2005, PP: 149:153.

Hala Elsadek and Dalia Nashaat, " Quad band Compact size Trapezoidal PIFA antenna", Journal of Electromagnetic Waves and applications (JEMWA), Vol. 21, No. 3, July 2007, pp. 865-876.

Hala Elsadek and Dalai Nashaat," Ultra Miniaturized E-shaped Dual band PIFA on cheap Foam and FR4 substrates", Journal of Electromagnetic Waves and Applications (JEMWA), Vol. 20, No.3, 2006, pp:291:300.

Hala Elsadek and Dalai Nashaat, " Multiband and UWB V-shaped Antenna Configuration for wireless Communications Applications", IEEE Antennas and Wireless Propagation Letters, Vol. 7, 2008, pp. 89-91.

Hirasawa and K. Fujimoto, "Characteristics of Wire Antenna on a Rectangular Conducting Body," lECE Trans., J65-B, 4, pp. 1133-1139, April 1982.

Hung-Jui Lam, Wideband Antenna in Coplanar Technology, M.Sc Thesis: University of Victoria, 2005

IAN OPPERMANN, LUCIAN STOICA, ALBERTO RABBACHIN, ZACK SHELBY,JUSSI HAAPOLA, "UWB Wireless Sensor Networks UWEN — A Practical Example", IEEE Radio Communications • December 2004

Jeen-Sheen Row "Dual-Frequency Triangular Planar Inverted-F antenna," IEEE Transactions on Antennas & Propagation, vol. 53, no. 2, pp. 874-876, February 2005.

J. Elling, M. Gentsch and J. Toftgrd, "Analysis of Integrated antennas Elsevier, 1990. for personal commuuicatioua," master's thesis, Aalborg University, Denmark, June 1991.

J. R. James and P. S. Hall, "Handbook of Microstrip Antennas," London, Peter Pereginus Ltd., 1989.

Kawala P. and U. H. Sheikh, "Adaptive multibeam array for wireless communications", Proc. Inst. Elect. Eng. 8th Int. Conf. Antennas and Propagation, Edinburgh, Scotland, 1993, pp: 970-974.

K. Fujimoto, A. Henderson, K. Hirasawa, and J. R. James, "Small Antennas," Research Studies Press LID., 1987.

K. Fujimoto and J. R. James, "Mobile Antenna Systems Handbook," Artech House, 1994.

K. Cho and Y. Yamada, "Impedance Characteristics of the Normal Mode Helical Antenna with a Nearby Conduction Plate," IEICE Trans., J73-8-II, 5, May 1990, pp. 250-256.

Ki-Hak Kim and Seong-Ook Park, "Analysis of the Small Band-Rejected Antenna with the parasitic strip for UWB", IEEE Transactions on Antennas & Propagation, Vol. 54, No 6, PP: 1688-1692, June 2006

K. P. Ray, "Design Aspects of Printed Monopole Antennas for Ultra-Wide Band Applications", Hindawi Publishing Corporation International Journal of Antennas and Propagation Volume 2008, Article ID 713858.

M. D. Yacoub, "Foundation of Mobile Radio Engineering," CRC Press, 1993.

Michael Chryssomallis," Smart antennas", IEEE Antenna and Propagation Magazine, vol. 42, No.3, 200, pp 129-136.

Min-Hua Ho, Ming-Ting Wu, Chung-I G. Hsu,and Jia-Yi Sze, "AN RHCP/LHCP SWITCHABLE SLOTLINE-FED SLOT-RING ANTENNA", MICROWAVE AND OPTICAL TECHNOLOGY LETTERS / Vol. 46, No. 1, July 5 2005

Mohamed H. Al Sharkawy, Abdelnasser A. Eldek, Atef Z. Elsherbeni and Charles E. Smith, "Design of Wideband Printed Monopole Antenna Using WIPL-D", April 19-23, 2004 - Syracuse, NY © 2004 ACES.

R. Al-Dahleh, C. Shafai, and L. Shafai, "FREQUENCY-AGILE MICROSTRIP PATCH ANTENNA USING A RECONFIGURABLE MEMS GROUND PLANE" MICROWAVE AND OPTICAL TECHNOLOGY LETTERS / Vol. 43, No. 1, October 5 2004.

Rappaport, T.S., "Smart antennas: Adaptive arrays, algorithms and wireless position location, New York, IEEE press, 1998.

R. Chair, K. M. Luk and K. F. Lee, "Small dual patch antenna," Electrons. Lett., vol. 35, pp. 762-763, May 1999.

S. I. Latif, L. Shafai, and S. K. Sharma, "Bandwidth enhancement and size reduction of microstrip slot antenna," IEEE Trans. Antennas Propag., vol. 53, pp. 994–1003, 2005.

Seok H. Choi, Jong K. Park,Sun K. Kim,and Jae Y. Park, "A NEW ULTRA-WIDEBAND ANTENNA FOR UWB APPLICATIONS", MICROWAVE AND OPTICAL TECHNOLOGY LETTERS / Vol. 40, No. 5, March 5 2004.

The international engineering consortium, "Smart antenna systems", Web proforum tutorials, 2005, http://www.iec.org

T.L.Roach, G. H. Huff and J. T. Bernhard," A comparative study of diversity gain and spatial coverage: fixed verses reconfigurable antennas for portable devices", Microwave and optical technology letters, vol. 49, No. 3, 2007, pp. 535-539.

Tsoulos G.V., "Adaptive antennas for wireless communications, IEEE press 2001.

T. Taga, "Analysis of planar inverted-F antennas and antenna design for portable radio equipment," in K. Hirasawa and M. Haneishi (Eds.), Analysis, Design and Measurement of small and Low-Profile Antennas, pp. 161-180, Norwood, MA, Artech House, 1992.

W.Stuzman, G.Theile, Antenna Theory and Design, 2nd edition, John Wiley & Sons, USA, 1998.

Yashar Zehforoosh, Changiz Ghobadi, and Javad Nourinia, "Antenna Design for Ultra Wideband Application Using a New Multilayer Structure", PIERS ONLINE, VOL. 2, NO. 6, 2006.

Yong Zhang, Bing-Zhong Wang, Xue-Song Yang and Weixia Wu, "A Fractal Hilbert Microstrip Antenna with Reconfigurable Radiation Patterns", Antennas and Propagation Society International Symposium, 2005 IEEE

Yujiang Wu and Zaiping Nie, "A novel reconfigurable microstrip antenna and its performance analysis in diversity reception system", Microwave and Optical Technology Letters, Vol. 49, No. 10, October 2007.

Large-Signal Modeling of GaN Devices for Designing High Power Amplifiers of Next Generation Wireless Communication Systems

Anwar Jarndal
Hodeidah University
Yemen

1. Introduction

An excellent candidate for fabrication of high-power amplifiers (HPAs) for next-generation wireless communication systems is a GaN HEMT. It has high sheet carrier density and high saturation electron velocity, which produce high output power. It also has high electron mobility, which is largely responsible for low on-resistance value which enhances high-power-added efficiency. As a result of GaN as a wideband material, the GaN HEMTs can achieve very high breakdown voltage and very high current density, and they can sustain very high channel-operating temperature. Furthermore, a possible epitaxial growth on silicon carbide substrate, which has excellent thermal properties, makes this device optimal for high-power RF applications. The past decade saw rapid progress in the development of GaN HEMTs with a focus on its power performance (Eastman et al., 2001). However, despite the high output power of this device, current dispersion is the biggest obstacle in obtaining reproducible power performance (Vetury et al., 2001); (Meneghesso et al., 2004). Designing an HPA based on the GaN HEMTs requires an accurate large-signal model for this device. This model should account for the current dispersion and temperature-dependent performance in addition to other high-power-stimulated effects like gate forward and breakdown phenomenon. In particular, the model should be able to predict intermodulation distortion (IMD), which is very important for the analysis of the HPA nonlinearity. In the last decade different models have been developed for GaN HEMTs. The analytical models reported in (Green et al., 2000) and (Lee & Webb, 2004) can simulate the fundamental output power including the current dispersion and thermal characteristics of the GaN HEMTs. However, these models have poor IMD-prediction capabilities. In another reported model (Raay et al., 2003), no IMD simulation has been presented. The model published in (Cabral et al., 2004) has been optimized for IMD simulation, but it does not account for the current dispersion or the temperature-dependent characteristics. This chapter addresses the development of a large-signal model for GaN HEMTs, which can simulate all of the mentioned effects in an efficient manner. First, a small-signal model that will be used as a basis for constructing the large-signal model will be described. Detailed steps for extraction of the small-signal model parameters will be presented. Large-signal

modeling and extraction procedures will also be explained. Finally, the developed large-signal model will be validated by comparing its simulations with measurements.

2. GaN HEMT

The general structure of the investigated devices is shown in Figure 1. The GaN HEMT structure was grown on SiC 2-inch wafers using Metal-Organic-Chemical-Vapour-Deposition (MOCVD) technology (Lossy[a] et al., 2002). This substrate provides an excellent thermal conductivity of 3.5 W/cm, which is an order of magnitude higher than that of sapphire. The epitaxial growth structure starts with the deposition of a 500 nm thick graded AlGaN layer on the substrate to reduce the number of threading dislocations in the GaN buffer layer due to the lattice mismatch between GaN and SiC layers. These threading dislocations enhance buffer traps and hence the associated drain-current dispersion (Hansen et al., 1998). A 2.7 µm thick highly insulating GaN buffer layer is then deposited to get lower background carrier concentration, which accordingly results in increased electron mobility in the above unintentionally doped layers. The buffer layer is followed by a 3 nm $Al_{0.25}Ga_{0.75}N$ spacer, 12 nm Si-doped $Al_{0.25}Ga_{0.75}N$ supply layer (5×10^{18} cm^{-3}), and 10 nm $Al_{0.25}Ga_{0.75}N$ barrier layer. The spontaneous and piezoelectric polarizations of the $Al_{0.25}Ga_{0.75}N$ layers form a two-dimensional electron gas (2DEG) at the AlGaN/GaN interface (Ambacher et al., 1999). The spacer layer is included to reduce the ionized-impurity scattering that deteriorates electron mobility in the 2DEG. The whole structure is capped with a 5 nm thick GaN layer to increase the effective Schottky barrier, which improves the breakdown characteristics and decreases the gate leakage. The measured 2DEG electron density and mobility, at room temperature, are 7.8×10^{12} cm^{-2} and 1400 cm^2/Vs (Lossy[a] et al., 2002). Device fabrication is accomplished using 0.5-µm stepper lithography, which results in an excellent homogeneity of the electrical properties over the wafer (Lossy et al., 2001).
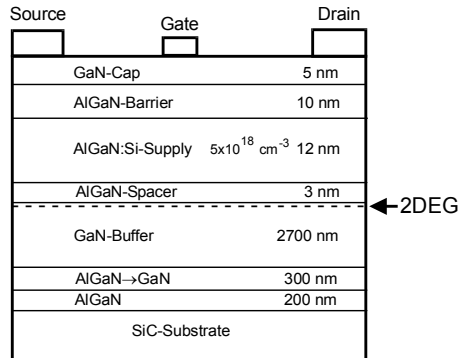
Fig. 1. Epitaxial layer structure of GaN HEMT.

Source and drain ohmic contacts have a metallization consisting of Ti/Al/Ti/Au/WSiN (10/50/25/30/120 nm) with improved edge and surface morphology. Due to the properties of the WSiN sputter deposition process, the Ti/Al/Ti/Au layers, which are deposited by e-beam evaporation, are totally embedded. The source and drain contacts are then rapidly thermal-annealed at 850 °C. The contact resistance is analyzed by Thermal Lens Microscope

(TLM) measurements with respect to thickness and composition of the different metallization layers at different temperatures. The contact resistance is determined to be 0.25-0.5 Ωmm under these conditions (Lossy[a] et al., 2002). Gate contacts are made from a Pt/Au metallization, and a gate length of 0.5μm is obtained using stepper lithography. Additionally, devices with gate length less than 0.3μm are written using a shaped electron beam tool (ZBA23-40kV) (Lossy[b] et al., 2002). SiN passivation layer is then deposited to reduce the surface trapping induced drain-current dispersion. Field plate connected to the gate, at the gate pad, and deposited over the passivation layer was employed for some investigated devices to improve its breakdown characteristics. An air-bridge technology using an electroplated Au is used to connect the source pads of multifinger devices.

3. Small-signal modeling

In bottom-up modeling technique, a multibias small-signal measurement is carried out over a range of bias points, and a large-signal model is then determined from the small-signal model derived at each of these bias points. Therefore, the accuracy of the constructed large-signal model depends on the accuracy of the bias-dependent small-signal model, which should reflect the electrical and physical characteristics of the device. Accurate determination of the intrinsic bias-dependent circuit of GaN HEMT small-signal model requires an efficient extraction method for the parasitic elements of the device. In (Jarndal & Kompa, 2005), an efficient reliable model parameter extraction method, applied for GaN HEMT, was developed. This method uses only a cold S-parameter measurement for accurate determination of the parasitic elements. The main advantage of this method is that it gives reliable values for the parasitic elements of the device without need for additional measurements or separate test patterns. Since the knowledge of distributed effects is important to identify the device parasitic elements for further minimization, a 22-element distributed model shown in Figure 2 is used as a small-signal model for GaN HEMT. This model is general and applicable for large gate periphery devices. The main advantages of this model are as follows.

 • It accounts for all expected parasitic elements of the device.
 • It reflects the physics of the device over a wide bias and frequency range.
Therefore, this model can be suitable for scalable large-signal model construction.
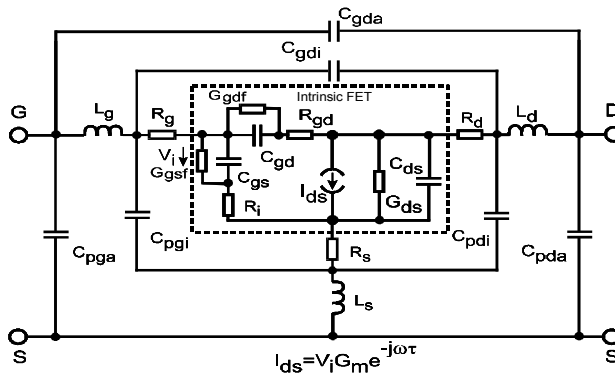
Fig. 2. 22-element distributed small-signal model for active GaN HEMT.
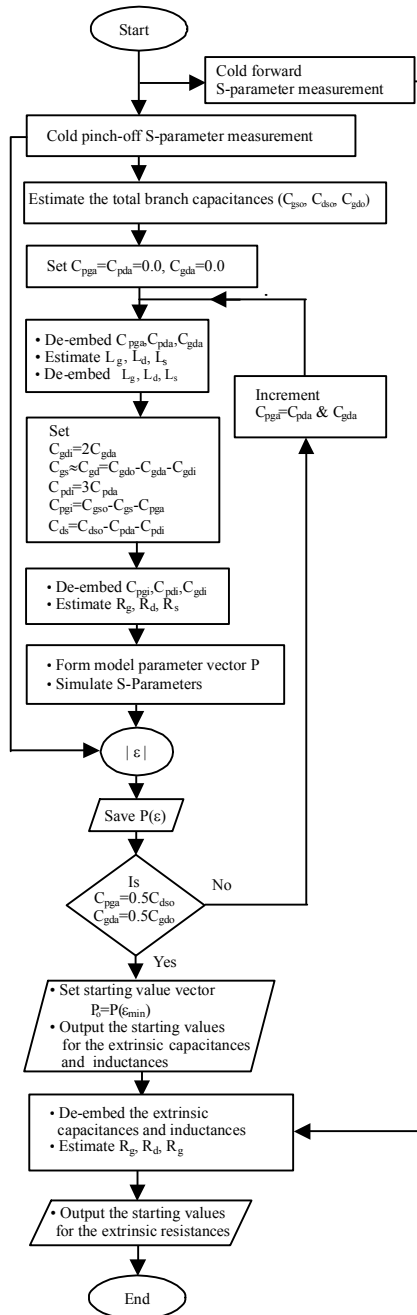
Fig. 3. Flowchart of the small-signal model parameter starting value generation algorithm.
© 2005 IEEE. Reprinted with permission.

In the extrinsic part of this model, C_{pga}, C_{pda} and C_{gda} account for parasitic elements due to the pad connections, measurement equipment, probes, and probe tip-to-device contact transitions; while C_{pgi}, C_{pdi}, and C_{gdi} account for interelectrode and crossover capacitances (due to air–bridge source connections) between gate, source, and drain. R_g, R_d, and R_s represent contact and semiconductor bulk resistances; while L_g, L_d, and L_s model effect of metallization inductances. In the intrinsic part, charging and discharging process for depletion region under the gate is described by C_{gs}, R_i, C_{gd} and R_{gd}. The gate forward and breakdown conductions are represented by G_{gsf} and G_{gdf}, respectively. Variation of the channel conduction with remote gate voltage is described by G_m; while the channel conductance controlled by local drain voltage is represented by G_{ds}. C_{ds} model the capacitance between the drain and source electrodes separated by the depletion region in electrostatic sense. Transit time of electrons in the channel at high-speed input signal is described by τ.

3.1 Extrinsic parameter extraction

Many of the model parameters in Figure 2 are difficult if not impossible to determine directly from measurements. Therefore, these parameters are determined through an optimization algorithm. The efficiency of this algorithm depends on the quality of starting values and the number of optimization variables. Under cold pinch-off condition, the equivalent circuit in Figure 2 can be simplified by excluding some elements, thereby reducing the number of unknowns. For further minimization of the number of optimization variables, only the extrinsic elements of the model will be optimized, while the intrinsic elements are determined from the deembedded Y-parameters. Under this bias condition, the reactive elements of the model are strongly correlated (Jarndal & Kompa, 2005). Therefore, the starting values estimation can be carried out in a way that takes this correlation into account. In addition, the S-parameter measurements must cover the frequency range where this correlation is more obvious. The required measurements frequency range for reliable starting values generation reduces for larger devices, e.g., up to 20 GHz for an 8x125-µm device. The proposed technique for starting values generation is based on searching for the optimal distribution of the total capacitances. This is achieved by scanning the outer capacitance values within the specified ranges. For each scanned value, the interelectrode capacitances are assigned suitable values and then deembedded from the measured Y-parameters. The rest of the model parameters are then estimated from the stripped Y-parameters. The whole estimated parameters are then used to simulate the device S-parameters, which are then compared with the measured ones. Using this systematic searching procedure, high-quality measurement-correlated starting values for the small-signal model parameters can be found. The closeness of the starting values to the real values simplifies the next step of parameters optimization since the risk of a local minimum is minimized.

A. Generation of starting value of small-signal model parameters

The starting values generation procedure is described by the flowchart in Figure 3. As shown in this flowchart, the starting values of the extrinsic capacitances and inductances are generated from pinch-off measurements, while those of extrinsic resistances are generated

from forward measurements. The whole starting values generating procedure can be summarized as follows.

Step 1)   Let $V_{GS} < -V_{pinch-off}$ and $V_{DS} = 0.0$ V. In this case, the equivalent circuit in Figure 2 of the active device can be used for this cold pinch-off device if the drain current source and the output channel conductance are excluded. Moreover, at low frequencies (in the megahertz range), this circuit can be reduced to a capacitive network shown in Figure 4 and the Y-parameters of this equivalent circuit can be written as

$$Y_{11} = j\omega \ (C_{gso} + C_{gdo})$$

(1)

$$Y_{22} = j\omega \ (C_{dso} + C_{gdo})$$

(2)

$$Y_{12} = Y_{21} = -j\omega \ C_{gdo}$$

(3)

where

$$C_{gdo} = C_{gda} + C_{gdi} + C_{gd}$$

(4)

$$C_{gso} = C_{pga} + C_{pgi} + C_{gs}$$

(5)

$$C_{dso} = C_{pda} + C_{pdi} + C_{ds} \ .$$

(6)
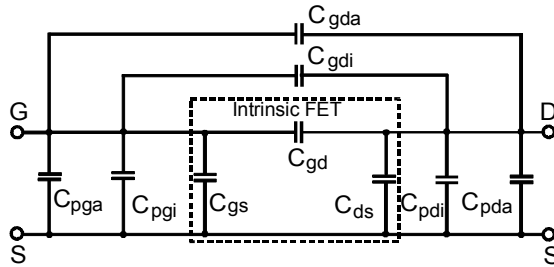
Fig. 4. Equivalent circuit model for cold pinch-off GaN HEMT at low frequency.

The total capacitances for gate–source, gate–drain, and drain–source branches are determined from the low frequency range of pinch-off S-parameter measurements, which are converted to Y-parameter.

Step 2)   The next step is searching for the optimal distribution of the total capacitances, which gives the minimum error between the measured and simulated S-parameters. This is achieved by scanning C_{pga}, C_{pda} , and C_{gda} values within the specified ranges. C_{pga} and C_{pda} are scanned from 0 to $0.5C_{dso}$ while C_{gda} is scanned from 0 to $0.5C_{gdo}$. During the scanning process, C_{pga} is assumed to be equal to C_{pda}

$$C_{pga} \cong C_{pda} \ .$$

(7)

The gate–drain interelectrode capacitance C_{gdi} is assumed to be twice the C_{gda} pad capacitance value

$$C_{gdi} \cong 2C_{gda} \ .$$

(8)

For symmetrical gate–source and gate–drain spacing, the depletion region will be uniform under pinch-off, so that

$$C_{gs} \cong C_{gd} = C_{gdo} - C_{gdi} - C_{gda} . \tag{9}$$

The value of C_{pgi} is calculated using

$$C_{pgi} = C_{gso} - C_{gs} - C_{pga} . \tag{10}$$

With the GaN devices under analysis, C_{pdi} is a significant part of the total drain–source capacitance. Therefore, it is found that the assumption

$$C_{pdi} \cong 3\,C_{pda} \tag{11}$$

minimizes the error between the simulated and measured S-parameters. For medium and high frequency range, the intrinsic transistor of the pinch-off model is represented in T-network as shown in Figure 5 where the interelectrode capacitances (C_{pgi}, C_{pdi} , and C_{gdi}) have been absorbed in the intrinsic capacitances (C_{gs} , C_{ds} , and C_{gd}). The values for C_{pga}, C_{pda} , and C_{gda} are deembedded from Y-parameter and then converted to Z-parameter.
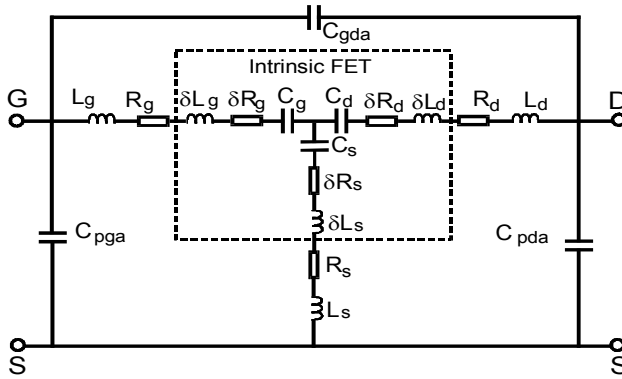
Fig. 5. Equivalent circuit model for cold pinch-off GaN HEMT at medium and high frequency.

This stripped Z-parameter can be written as

$$Z_{11} = R_g + R_s + j\omega(L_g + L_s) + \frac{1}{j\omega}\left(\frac{1}{C_g} + \frac{1}{C_s}\right) + \delta Z_g \tag{12}$$

$$Z_{22} = R_d + R_s + j\omega(L_d + L_s) + \frac{1}{j\omega}\left(\frac{1}{C_d} + \frac{1}{C_s}\right) + \delta Z_d \tag{13}$$

$$Z_{12} = Z_{21} = R_s + j\omega L_s + \frac{1}{j\omega C_s} + \delta Z_s \tag{14}$$

where

$$\delta Z_g = \delta R_g + \delta R_s + j\omega(\delta L_g + \delta L_s) \tag{15}$$

$$\delta Z_d = \delta R_d + \delta R_s + j\omega(\delta L_d + \delta L_s) \tag{16}$$

$$\delta Z_s = \delta R_s + j\omega \delta L_s .$$ (17)

δZ_g, δZ_d, and δZ_s represent correction terms related to the intrinsic parameters of the model. Ignoring the correction terms and multiplying the Z-parameters by ω and then taking the imaginary parts gives

$$\mathrm{Im}[\omega Z_{11}] = (L_g + L_s)\omega^2 - \left(\frac{1}{C_g} + \frac{1}{C_s}\right)$$ (18)

$$\mathrm{Im}[\omega Z_{22}] = (L_d + L_s)\omega^2 - \left(\frac{1}{C_d} + \frac{1}{C_s}\right)$$ (19)

$$\mathrm{Im}[\omega Z_{12}] = L_s\omega^2 - \frac{1}{C_s} .$$ (20)

Hence, the values of L_g, L_d , and L_s can be extracted from the slope of $\mathrm{Im}[\omega Z_{ij}]$ versus ω^2 curve. The estimated values of the inductances described above and the interelectrode capacitances (C_{pgi}, C_{pdi}, and C_{gdi}) are deembedded. However, the incomplete deembedding of the outer capacitances and the inductances introduce nonlinear frequency dependence in the real part of deembedded Z-parameters. By multiplying the deembedded Z-parameter by ω^2, this effect is reduced (Jarndal & Kompa, 2005). Ignoring the correction terms and multiplying the deembedded Z-parameter by ω^2 and then taking the real part of this Z–parameter gives

$$\mathrm{Re}[\omega^2 Z_{11}] = \omega^2 (R_g + R_s)$$ (21)

$$\mathrm{Re}[\omega^2 Z_{22}] = \omega^2 (R_d + R_s)$$ (22)

$$\mathrm{Re}[\omega^2 Z_{12}] = \omega^2 R_s .$$ (23)

By linear regression, the value of R_g+R_s, R_d+R_s, and R_s can be extracted from the slope of $\mathrm{Re}[\omega^2 Z_{ij}]$ versus ω^2 curves. The resulting estimated parameters are used to simulate the device S-parameters, which are then compared with the measured ones to calculate the residual fitting error (ε). The outer capacitances (C_{pga}, C_{pda}, and C_{gda}) are incremented, and the procedure is repeated until C_{pga} (C_{pda}) is equal to $0.5C_{dso}$ and C_{gda} is equal to $0.5C_{gdo}$. The vector of model parameters $P(\varepsilon_{min})$, corresponding to the lowest error ε_{min}, is then taken as the appropriate starting value.

Step 3)   Because of unavoidable high measurement uncertainty for cold pinch-off device, the determination of a reliable starting value for the extrinsic resistances is difficult if not impossible. More reliable starting value was generated using cold gate forward S-parameter measurements at high gate voltage greater than or equal to 2 V. This is due to the higher conduction band of GaN-based HEMT with respect to the corresponding GaAs-based HEMT. Therefore, significantly higher voltages have to be applied to reach the condition when the influence of the gate capacitance is negligible. The determined values of extrinsic capacitances and inductances, in Step 2), are deembedded from the gate-forward measurements. The starting values of the extrinsic resistances are then estimated from the stripped forward Z–parameters.

B. Model parameter optimization

The procedure for the generation of starting values of the model parameters was discussed in Section 3.1-A. Here, the result of the optimal value for each model parameter is presented. Model parameters optimization has been done based on the principle of bidirectional optimization technique proposed by Lin and Kompa (Lin & Kompa, 1994). This technique works successfully for lumped small-signal model, but cannot be used efficiently for distributed model. This is due to the extra parasitic elements of this model, which increase the searching space. Now, this algorithm can be modified to become applicable for the distributed model, where the closeness of the generated starting value to the true value allows the searching space to be reduced by optimizing only the extrinsic parameters. At each iteration through the optimization process, the extrinsic parameters are assigned suitable values and then deembedded from the measured data to determine the intrinsic Y-parameters. The intrinsic model parameters are then estimated by means of data fitting from the deembedded measurements. The whole estimated model parameters are then used to fit the measured S-parameters. This process is continued to find the optimal model parameters. In this case, the optimization problem is a nonlinear multidimensional one, whose objective function is likely to have multiple local minima. Furthermore, the cold pinch-off device measurements have a high uncertainty (HP8510B network analyzer manual, 1987). These two factors increase the probability of trapping into a local minimum, which requires a careful formulation for the objective function to avoid the local minimum problem. The magnitude of the error between the measured S-parameter and its simulated value can be expressed as (Jarndal & Kompa, 2005)

$$\varepsilon_{ij} = \frac{\left|\mathrm{Re}(\delta S_{ij,n})\right| + \left|\mathrm{Im}(\delta S_{ij,n})\right|}{W_{ij}} \qquad , \mathrm{i,j} = 1,2; \ n = 1,2,\ldots,N \tag{24}$$

where

$$W_{ij} = \max[\left|S_{ij}\right|] \qquad , \mathrm{i,j} = 1,2; \ \mathrm{i} \neq \mathrm{j} \tag{25}$$

$$W_{ii} = 1 + \left|S_{ii}\right| \qquad , \mathrm{i} = 1,2 \tag{26}$$

and N is the total number of data points. δS is the difference between the measured S-parameter coefficient and its simulated value. The weighting factor (W) deemphasizes data region with higher reflection coefficients due to the involved higher measurement uncertainty. The scalar error is then expressed as

$$\varepsilon_s = \frac{1}{N} \sum_{n=1}^{N} \left\| \overset{*}{\varepsilon}(f_n) \right\|_1 \tag{27}$$

where

$$\overset{*}{\varepsilon}(f_n) = \begin{bmatrix} \varepsilon_{11}(f_n) & \varepsilon_{12}(f_n) \\ \varepsilon_{21}(f_n) & \varepsilon_{22}(f_n) \end{bmatrix} \tag{28}$$

defined at each frequency point. However, the objective function that is based on S-parameters alone to minimize the fitting error may not necessarily lead to physically relevant values of the model parameters (Kompa & Novotny, 1997). For further

enhancement of the objective function, another performance quantity, depending on the final application, will be considered. The main application of GaN-based HEMT is power amplifier design. For power amplifier design, the output and input impedance, the device gain, and stability factor are important for the design of matching networks. These factors can be expressed as a function of S–parameters and fitted during the optimization. The stability factor defined at the output plane of the device at each frequency can be expressed as

$$K = \frac{1 - \left|S_{22}\right|^2}{\left|S_{22} - S_{11}^* \Delta_s\right| + \left|S_{12}S_{21}\right|}$$

(29)

where S^* is the complex conjugate and Δ_s is the determinant of S-parameter matrix at each frequency (Edwards & Sinsky, 1992). The fitting error of the stability factor is given by

$$\varepsilon_K = \frac{1}{N}\sum_{k=1}^{N}\left|K_{meas} - K_{sim}\right|$$

(30)

where K_{meas} and K_{sim} are the stability factors from the measured and simulated S-parameters, respectively. With regard to the device gain, the maximally efficient gain defined in (Kotzebue,1976) is a more suitable one, since it remains finite even for an unstable device. This gain may be defined at each frequency as

$$G = \frac{\left|S_{21}\right|^2 - 1}{\ln\left|S_{21}\right|^2}.$$

(31)

The error in the gain may thus, be expressed as

$$\varepsilon_G = \frac{1}{N}\sum_{m=1}^{N}\left|G_{meas} - G_{sim}\right|$$

(32)

where G_{meas} and G_{sim} are the gains computed from the measured and modeled S-parameters. The fitting error can be defined in terms of the three error components as

$$\varepsilon = \sqrt{\frac{1}{3}\left(\varepsilon_s^{\ 2} + \varepsilon_K^{\ 2} + \varepsilon_G^{\ 2}\right)}.$$

(33)

The modified Simplex optimization algorithm proposed in (Kompa & Novotny, 1997) is used to minimize the objective function in (33).

The extraction procedure was applied to different GaN HEMT sizes. Table 1 presents the final optimised results for extrinsic parameters extraction. As it can be observed in the table, the extracted pad capacitances (C_{pga}, C_{pda}, and C_{gda}) are in proportion with the gate width. There is no significant difference between the pad capacitances of 8x125-µm and 8x250-µm devices because the pad connection area is related mainly to the number of fingers. The inter-electrode capacitances (C_{pdi} and C_{gdi}) are also in proportion with the gate width. Due to the small values of R_g and R_s, for larger devices, C_{pgi} cannot be separated completely from the intrinsic capacitance C_{gs}. However, the sum of C_{pgi} and C_{gs} is in proportion with the gate width. By direct scaling of the 8x250-µm device, the expected values of C_{gda} and C_{gdi} for

8x125-µm device are 20 fF and 40 fF, respectively. Due to the smaller values of these elements and also due to the smaller values of L_g and L_d for this device, C_{gda} and C_{gdi} cannot be separated form C_{gd}. The parasitic inductance includes the self-inductance due the metallization contact and the mutual inductance between the metal interconnection. The mutual inductance increases by increasing the number of fingers. For this reason, there is a considerable increase of L_d and L_g values for 16x250-µm device with respect to 8x125-µm device (Jarndal[a] & Kompa, 2006). The parasitic resistances (R_d and R_s) are inversely proportional with the gate width. However, this is not the case with R_g, which is proportional with the unite-gate-width and inversely proportional with the number of gate fingers as reported in (Goyal et al., 1989).

Parameter	W_g = 16x250 µm	W_g = 8x250 µm	W_g = 8x125 µm	W_g = 2x50 µm
C_{pga} (fF)	233.5	89.8	86.9	9.97
C_{pgi} (fF)	39.6	234.8	332.2	7.09
C_{gs} (fF)	1508.4	538.6	255.8	15.38
C_{gda} (fF)	121.6	41.7	0.0	0.47
C_{gdi} (fF)	265.6	96.5	0.0	0.86
C_{gd} (fF)	1285.7	757.8	517.4	20.17
C_{pda} (fF)	206.4	90.9	86.3	7.13
C_{pdi} (fF)	790.7	390.2	245	29.42
C_{ds} (fF)	0.0	0.0	1.0	0.0
L_g (pH)	122.3	81.9	57.3	46.55
L_d (pH)	110.9	75.4	54.5	47.9
L_s (pH)	3.6	5.7	5.6	6.25
R_g (Ω)	1.1241	2.8	1.7	4.8
R_d (Ω)	0.71424	1.4	2.3	11.8
R_s (Ω)	0.25152	0.5	0.9	5.47
R_i (Ω)	0.0	0.0	0.0	0.0
R_{gd} (Ω)	0.1	0.0	0.0	0.0
G_m (mS)	0.0	0.0	0.0	0.0
τ (ps)	0.0	3.3	0.0	0.0
G_{ds} (mS)	0.34	0.0	0.26	0.0
G_{gsf} (mS)	2.3	0.6	0.4	0.0
G_{gdf} (mS)	0.24	0.25	0.2	0.0

Table 1. Extracted model parameters for different GaN HEMT sizes under cold pinch-off bias condition (V_{DS} = 0 V and V_{GS} = $V_{pinch-off}$). © 2006 IEEE. Reprinted with permission.

3.2 Intrinsic parameter extraction

After deembedding the extracted extrinsic parameters in Section 3.1, the bias-dependent intrinsic parameters can be extracted. An efficient technique is developed for extracting of the optimal value of the intrinsic element. In this technique, the intrinsic Y–parameters are formulated in a way where the optimal intrinsic element value can be extracted using simple linear data fitting (Jarndal & Kompa, 2005). The admittance for the intrinsic gate–source branch Y_{gs} is given by

$$Y_{gs} = Y_{i,11} + Y_{i,12} = \frac{G_{gsf} + j\omega C_{gs}}{1 + R_i G_{gsf} + j\omega R_i C_{gs}}. \tag{34}$$

By defining a new variable D as

$$D = \frac{|Y_{gs}|^2}{Im[Y_{gs}]} = \frac{G_{gsf}{}^2}{\omega C_{gs}} + \omega C_{gs}. \tag{35}$$

C_{gs} can be determined from the slope of the curve for ωD versus ω^2 by linear fitting, where ω is the angular frequency. By redefining D as

$$D = \frac{Y_{gs}}{Im[Y_{gs}]} = \frac{G_{gsf}(1 + R_i G_{gsf})}{\omega C_{gs}} + \omega R_i C_{gs} - j. \tag{36}$$

R_i can be determined from the plot of the real part of ωD versus ω^2 by linear fitting. G_{gsf} can be determined from the real part of Y_{gs} at low frequencies (in the megahertz range). The admittance for the intrinsic gate–drain branch Y_{gd} is given by

$$Y_{gd} = -Y_{i,12} = \frac{G_{gdf} + j\omega C_{gd}}{1 + R_{gd} G_{gdf} + j\omega R_{gd} C_{gd}}. \tag{37}$$

The same procedure, given in (35) and (36), can be used for extracting C_{gd}, R_{gd}, and G_{gdf}. The admittance of the intrinsic transconductance branch Y_{gm} can be expressed as

$$Y_{gm} = Y_{i,21} - Y_{i,12} = \frac{G_m e^{-j\omega\tau}}{1 + R_i G_{gsf} + j\omega C_{gs}}. \tag{38}$$

By redefining D as

$$D = \left|\frac{Y_{gs}}{Y_{gm}}\right|^2 = \left(\frac{G_{gsf}}{G_m}\right)^2 + \left(\frac{C_{gs}}{G_m}\right)^2 \omega^2. \tag{39}$$

G_m can be determined from the slope of the curve for D versus ω^2 by linear fitting. By redefining D as

$$D = (G_{gsf} + j\omega C_{gs})\frac{Y_{gm}}{Y_{gs}} = G_m e^{-j\omega\tau}. \tag{40}$$
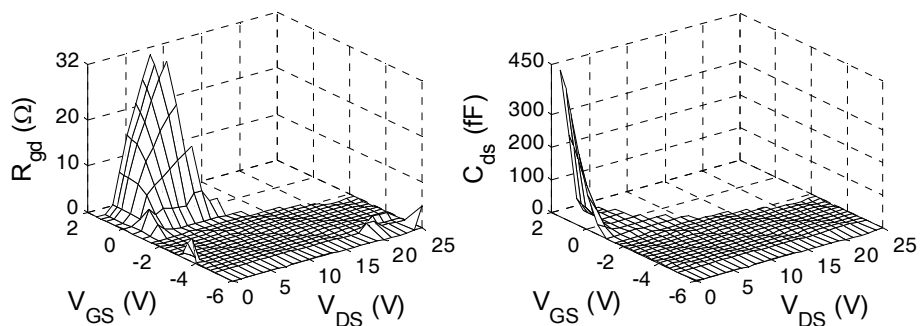
Fig. 6. Extracted R_{gd} and C_{ds} as a function of the extrinsic voltages for a GaN HEMT with a 2x50-µm gate width. © 2005 IEEE. Reprinted with permission.
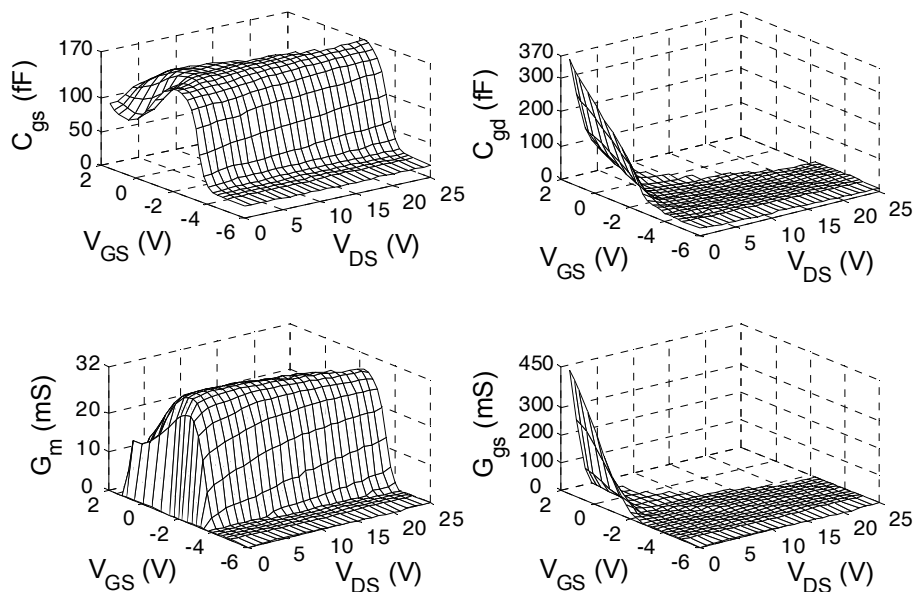
Fig. 7. Extracted C_{gs}, C_{gd}, G_m, and G_{ds} as a function of the extrinsic voltages for a GaN HEMT with a 2x50-µm gate width. © 2005 IEEE. Reprinted with permission.
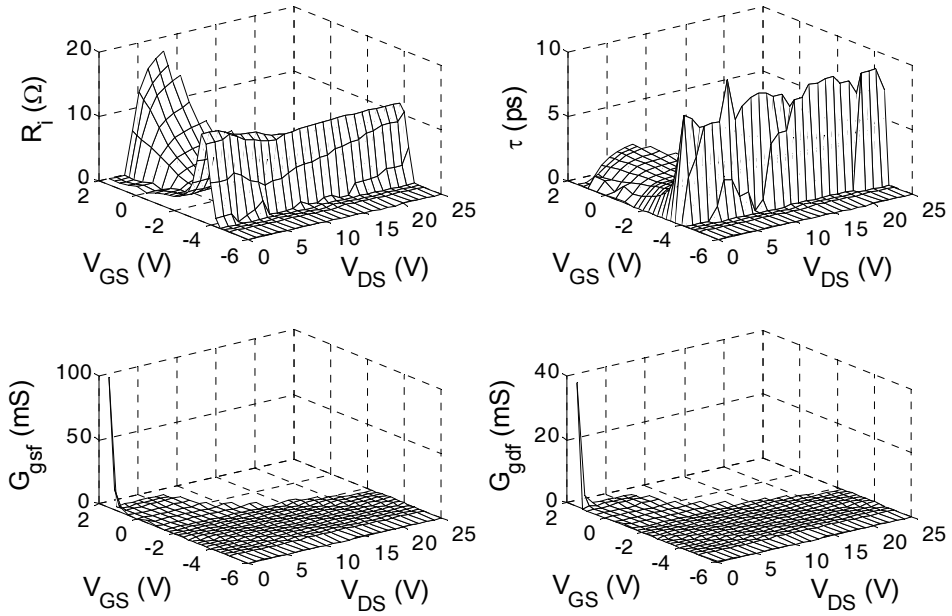
Fig. 8. Extracted R_i, τ, G_{gsf}, and G_{gdf} as a function of the extrinsic voltages for a GaN HEMT with a 2x50-µm gate width. © 2005 IEEE. Reprinted with permission.

τ can be determined from the plot of the phase of D versus ω by linear fitting. The admittance of the intrinsic drain–source branch Y_{ds} can be expressed as

$$Y_{ds} = Y_{i,22} + Y_{i,12} = G_{ds} + j\omega C_{ds} \,. \tag{41}$$

C_{ds} can be extracted from the plot of the imaginary part of Y_{ds} versus ω by linear fitting. Due to the frequency-dependent effect in the output conductance G_{ds}, its value is determined from the curve of $\omega \mathrm{Re}[Y_{ds}]$ versus ω by linear fitting.

Figs. 6-8 present extracted intrinsic parameters for GaN HEMT using the proposed procedure under different extrinsic bias voltages. The extraction results show the typical expected characteristics of GaN HEMT. The reliability of the extraction results was demonstrated in (Jarndal & Kompa, 2005) in terms of the reverse modeling of the effective gate length for the same analysed devices. The accuracy of the proposed small signal modeling approach is verified through S-parameter simulation for different device sizes under different bias conditions. As it can be seen in Figs. 9 and 10, the model can simulate the S-parameter accurately. Also it can predict the kink effect in S_{22}, which occurs in larger size FETs (Lu et al., 2001).
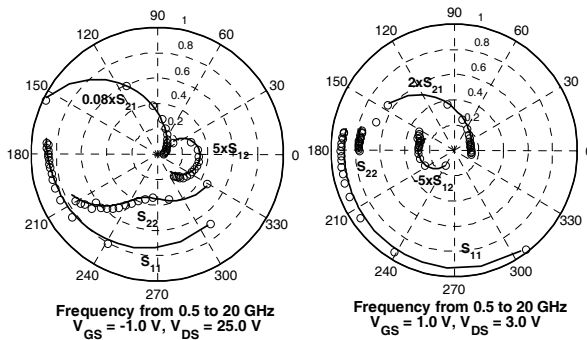
Fig. 9. Comparison of measured S-parameters of a 8x125-µm GaN HEMT (circles) with simulation results (lines) at (V_{GS} = -1, V_{DS} = 25 V) and (V_{GS} = 1 V, V_{DS} = 3 V). © 2006 IEEE. Reprinted with permission.
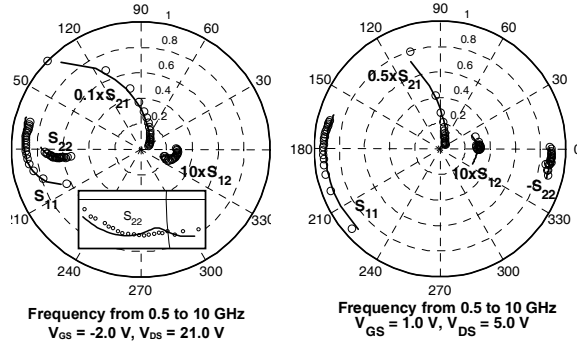
Fig. 10. Comparison of measured S-parameters of a 16x250-µm GaN HEMT (circles) with simulation results (lines) at (V_{GS} = -2, V_{DS} = 21 V) and (V_{GS} = 1 V, V_{DS} = 5 V). © 2006 IEEE. Reprinted with permission.

4. Large-signal modeling

Under RF large-signal operation, the values of the intrinsic-elements of the GaN HEMT model in Figure 2 vary with time and become dependent on the terminal voltages. Therefore the intrinsic part of this model can be described by the equivalent-circuit model shown in Figure 11. In this circuit, two quasi-static gate-current sources I_{gs} and I_{gd} and two quasi-static gate-charge sources Q_{gs} and Q_{gd} are used to describe the conduction and displacement currents. The nonquasi-static effect in the channel charge is approximately modeled with two bias-dependent resistors R_i and R_{gd} in series with Q_{gs} and Q_{gd}, respectively. This implementation is simpler and it improves the accuracy of the model up to millimeter-wave frequencies (Schmale & Kompa, 1997). A nonquasistatic drain-current model which accounts for trapping and self-heating effects is embedded in the proposed large-signal model. The drain-current value is determined by the applied intrinsic voltages V_{gs} and V_{ds}, whereas the amount of trapping induced current dispersion is controlled by the ac components of these voltages. These components are extracted from the intrinsic voltage

using RC high-pass circuits at gate and drain sides, as shown in Figure 11. The capacitors C_{GT} and C_{DT} values are selected to be 1 pF to provide a "macroscopic" modeling of charges stored in the surface and buffer traps. These charges are almost related to the leakage currents from the gate metal edge to the surface (Vetury et al., 2001) or from the channel into the buffer layer (Kohn et al., 2003). The small leakage currents in the gate and drain paths are realized with large (on the order of 1MΩ) resistances R_{GT} and R_{DT} in series with C_{GT} and C_{DT}, respectively.
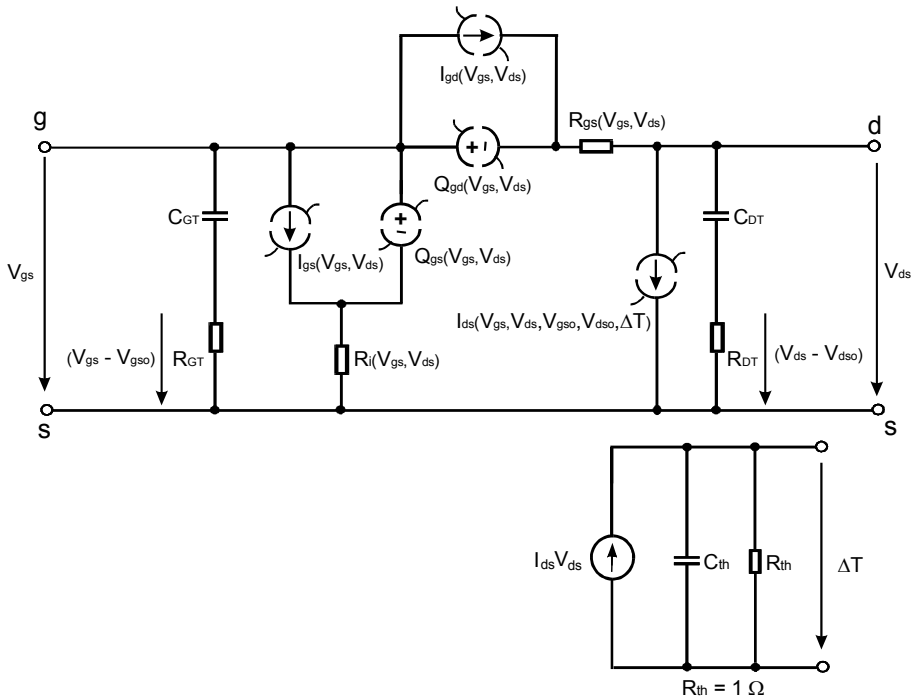
Fig. 11. Large-signal model for GaN HEMT including self-heating and trapping effects.

This implementation makes the equivalent circuit more physically meaningful; moreover, it improves the model accuracy for describing the low-frequency dispersion, as shown in Figure 12. This figure shows simulated frequency dispersion of the channel transconductance and output conductance, which is related mainly to the surface and buffer traps. The values of R_{GT}, R_{DT}, C_{GT}, and C_{DT} are chosen to result in trapping time constants on the order of $10^{-5} - 10^{-4}$ s (Meneghesso et al., 2001). In the current model, the amount of self-heating-induced current dispersion is controlled by normalized channel temperature rise ΔT. The normalized temperature rise is the channel temperature divided by the device thermal resistance R_{th}. A low-pass circuit is added to determine the value of ΔT due to the static and quasi-static dissipated power. The value of the thermal capacitance C_{th} is selected to define a transit time constant on the order of 1 ms (Kohn et al., 2003). R_{th} is normalized to one because its value is incorporated in thermal fitting parameter in the current-model expression, as will be discussed in section 4.2.
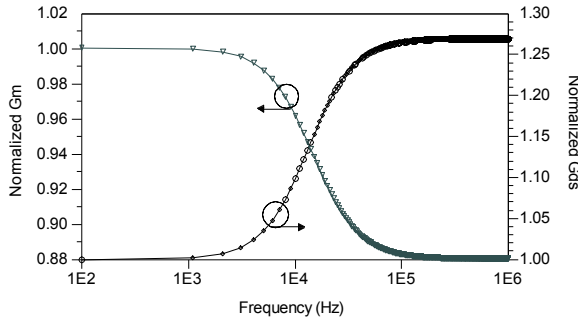
Fig. 12. Simulated normalized transconductance and output conductance for a 8x125-µm GaN HEMT at V_{DS} = 24 V and V_{GS} = -2 V.

4.1. Gate charge and current modeling

The intrinsic elements are extracted as a function of the extrinsic voltages V_{GS} and V_{DS} as presented in Figs. 6-8 for 2x50-µm GaN HEMT. To determine the intrinsic charge and current sources of the large-signal model by integration, a correction has to be carried out that considers the voltage drop across the extrinsic resistances. Therefore, the intrinsic voltages can be calculated as

$$V_{ds} = V_{DS} - (R_d + R_s)I_{ds} - R_s I_{gs} \tag{42}$$

$$V_{gs} = V_{GS} - (R_g + R_s)I_{gs} - R_s I_{ds}. \tag{43}$$

This implies that the values of the intrinsic voltages V_{gs} and V_{ds} are no longer equidistant, which makes the intrinsic-element integration difficult to achieve. In addition, this representation is not convenient to handle in Advanced Design System (ADS) simulator. Interpolation technique can be used to uniformly redistribute the intrinsic element data with respect to equidistant intrinsic voltages. However, the main limitations of this technique are that it produces discontinuities and an almost oscillating behavior in the interpolated data. These effects result in inaccurate simulation of higher order derivatives of the current and charge sources, which deteriorate output-power harmonics and IMD simulations (Cuoco et al., 2002). Therefore, B-spline-approximation technique is used for providing a uniform data for the intrinsic elements (Jarndal & Kompa, 2007). This technique can maintain the continuity of the data and its higher derivatives and hence improves the model simulation for the harmonics and the IMD (Koh et al., 2002). Generally, the intrinsic gate capacitances and conductances satisfy the integration path-independence rule (Root et al., 1991). Thus, the gate charges can be determined by integrating the intrinsic capacitances C_{gs}, C_{gd}, and C_{ds} as follows (Schmale & Kompa, 1997):

$$Q_{gs}(V_{gs}, V_{ds}) = \int_{V_{gs0}}^{V_{gs}} C_{gs}(V, V_{ds0})\, dV + \int_{V_{ds0}}^{V_{ds}} C_{ds}(V_{gs}, V)\, dV \tag{44}$$

$$Q_{gd}(V_{gs},V_{ds}) = \int_{V_{gs0}}^{V_{gs}} C_{gd}(V,V_{dso})\,dV$$
$$- \int_{V_{ds0}}^{V_{ds}} [C_{ds}(V_{gs},V) + C_{gd}(V_{gs},V)]\,dV$$

(45)

where V_{gs0} and V_{ds0} are arbitrary starting points for the integration. The shapes of the calculated Q_{gs} and Q_{gd}, shown in Figure 13, for GaN HEMTs are similar to the reported ones for AlGaAs/GaAs HEMTs in (Schmale & Kompa, 1997). The gate currents I_{gs} and I_{gd} are determined by the integration of the intrinsic gate conductances G_{gfs} and G_{gdf} as follows:

$$I_{gs}(V_{gs},V_{ds}) = I_{gs}(V_{gs0},V_{ds0}) + \int_{V_{gs0}}^{V_{gs}} G_{gsf}(V,V_{ds0})\,dV$$

(46)

$$I_{gd}(V_{gs},V_{ds}) = I_{gd}(V_{gs0},V_{ds0}) + \int_{V_{gs0}}^{V_{gs}} G_{gdf}(V,V_{ds0})\,dV$$
$$- \int_{V_{ds0}}^{V_{ds}} G_{gdf}(V_{gs},V)\,dV$$

(47)

The calculated values of I_{gs} and I_{gd} as a function of the intrinsic voltages are illustrated in Figure 14.
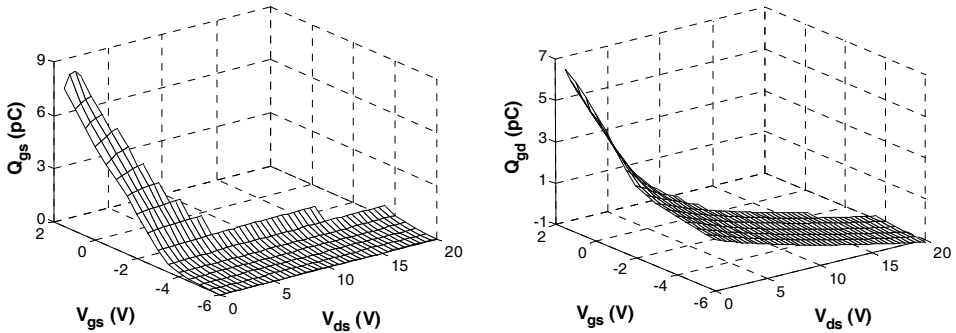
Fig. 13. Calculated gate-charge sources Q_{gs} and Q_{gd} versus intrinsic voltages for a 8x125-µm GaN HEMT. © 2007 IEEE. Reprinted with permission.
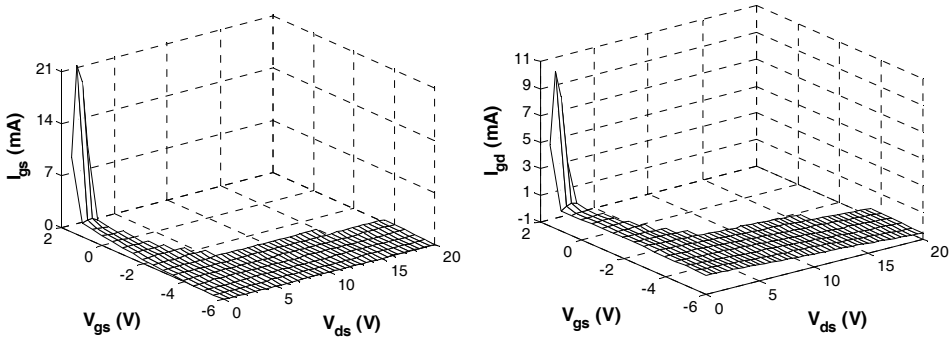
Fig. 14. Calculated gate-current sources I_{gs} and I_{gd} versus intrinsic voltages for a 8x125-µm GaN HEMT. © 2007 IEEE. Reprinted with permission.

4.2. Drain–current modeling

Due to self-heating and trapping effects, associated with high-power devices, the intrinsic channel conductance and transconductance (G_{ds} and G_m) do not satisfy the integration path-independence rule (Wei et al., 1999). Therefore, the RF drain current cannot be derived by relying on conventional S-parameter measurements. In addition, the self-heating and trapping cannot be characterized separately by these measurements to get an accurate current model. The optimal method is to derive the current model from pulsed I–V measurements under appropriate quiescent bias conditions, as presented in (Jarndal[b] et al., 2006). The drain current is modeled as (Filicori et al., 1995)

$$
\begin{aligned}
I_{ds}(V_{ds}, V_{gs}, V_{dso}, V_{gso}, P_{diss}) = {} & I_{ds,iso}^{DC}(V_{gs}, V_{ds}) \\
& + \alpha_G(V_{gs}, V_{ds})(V_{gs} - V_{gso}) \\
& + \alpha_D(V_{gs}, V_{ds})(V_{ds} - V_{dso}) \\
& + \alpha_T(V_{gs}, V_{ds})P_{diss}
\end{aligned}
\tag{48}
$$

where $I^{DC}_{ds,iso}$ is the isothermal dc current after deembedding the self-heating effect. α_G and α_D model the deviation in the drain current due to the surface-trapping and buffer-trapping effects, respectively, and α_T models the deviation in the drain current due to the self-heating effect. The amount of trapping-induced current dispersion depends on the rate of dynamic change of the applied intrinsic voltages V_{gs} and V_{ds} with respect to those average values V_{gso} and V_{dso}. In other words, this current dispersion is mainly stimulated by the RF or the ac components of the gate–source and drain–source voltages, which is described by ($V_{gs} - V_{gso}$) and ($V_{ds} - V_{dso}$) in (48). The self-heating-induced dispersion is caused mainly by the low-frequency components of the drain signal. Therefore, P_{diss} in (48) accounts for the static and quasistatic intrinsic power dissipation.

A. Trapping and self-heating characterization

Trapping effects can be characterized by pulsed I–V measurements at negligible device self-heating (Charbonniaud et al., 2003). The surface trapping is characterized by pulsed I–V's at two extrinsic quiescent biases equivalent to:

$$V_{GSO} < V_P, \; V_{DSO} = 0 \; V \; (P_{diss} \approx 0)$$

$$V_{GSO} = 0 \; V, \; V_{DSO} = 0 \; V \; (P_{diss} \approx 0).$$

The buffer trapping is characterized by pulsed I–V 's at two quiescent biases equivalent to:

$$V_{GSO} < V_P, \; V_{DSO} = 0 \; V \; (P_{diss} \approx 0)$$

$$V_{GSO} < V_P, \; V_{DSO} >> 0 \; V \; (P_{diss} \approx 0).$$

These two conditions lead to different states of the trapping effects but involve negligible power dissipation. To characterize the self-heating, additional pulsed I–V characteristics at rather high quiescent power dissipation are used. DC I–V characteristics can also be used in addition to the pulsed I–V characteristics for further improvement of the self-heating characterization (Jarndal[b] et al., 2006).

B. Drain–current-model parameter extraction
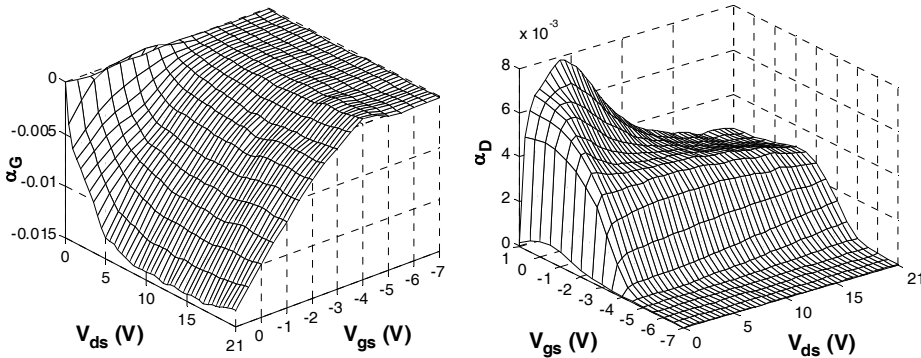
Fig. 15. Bias-dependent trapping fitting parameters of the drain–current model in (48) extracted from the pulsed I–V measurements of a 8x125-µm GaN HEMT. © 2007 IEEE. Reprinted with permission.

The drain–current-model equation in (48) has four unknowns: $I^{DC}_{ds,iso}$, a_G, a_D, and a_T. To determine these unknowns, the equation should be applied to, at least, four pulsed I–V characteristics at suitable quiescent bias conditions that lead to four highly independent linear equations. The described I–V characteristics in Section 4.2-A define approximately four independent states for the drain current. At each state, the drain current can be assumed to be affected by, at most, one of the dispersion sources (surface trapping, buffer trapping, or self-heating). By solving the four linear equations, corresponding to the four characteristics, at each bias point, the values of $I^{DC}_{ds,iso}$, a_G, a_D, and a_T can be determined. Figs. 15 and 16 show the extracted values of these fitting parameters as a function of the intrinsic voltages.

(a)   (b)
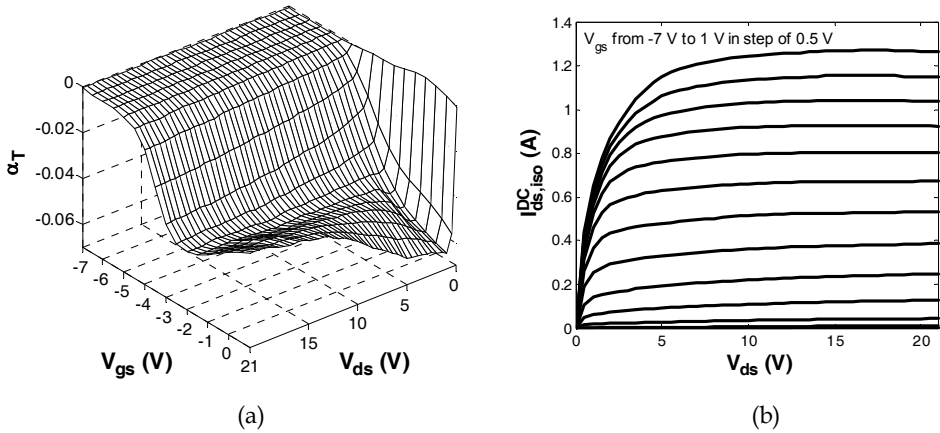
Fig. 16. (a) Extracted bias-dependent self-heating fitting parameter and (b) isothermal dc drain current for a 8x125-µm GaN HEMT. © 2007 IEEE. Reprinted with permission.

4.3 Large-signal model implementation and verification
The large-signal model was implemented as a table-based model in ADS. The extrinsic bias-independent passive elements are represented by lumped elements, whereas the intrinsic nonlinear part is represented by a symbolically defined device (SDD) component.
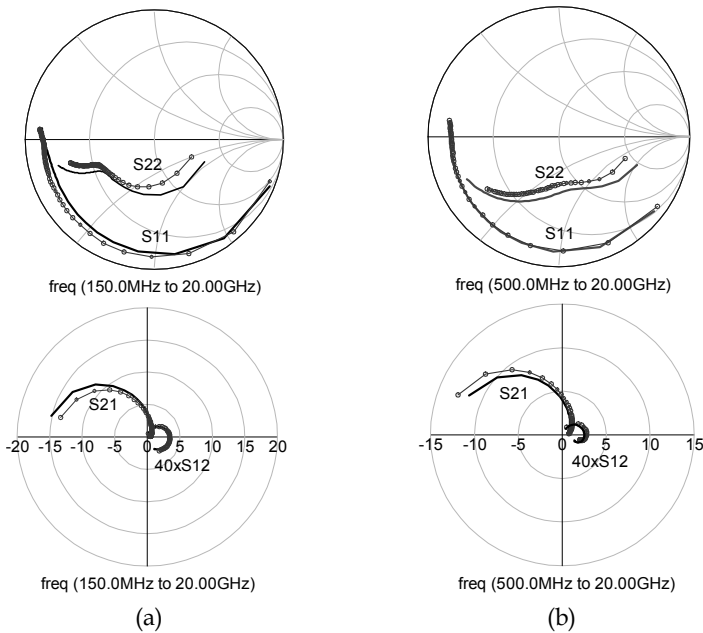

(a)   (b)

Fig. 17. (Lines) Simulated and (circles) measured S-parameters of a 8x125-µm GaN HEMT at (a) V_{GS} = −2.0 V and V_{DS} = 9.0 V and (b) V_{GS} = −3.0 V and V_{DS} = 21.0 V.

The developed large-signal model was verified by independent measurements. The considered devices are 8×125-µm GaN HEMTs on different wafers. First, the model is checked whether it is consistent with I–V and S-parameter measurements it has been derived from. Second, large-signal single- and two-tone simulations are compared with measurements. S-parameter simulation in comparison with measurement of a 8×125-µm device is shown in Figure 17. The good agreement between simulation and measurement verifies the consistency of the large-signal model with the small-signal equivalent-circuit model. Pulsed I–V simulation has been done at quiescent bias conditions different than the used ones for model parameter extraction. Figure 18 shows pulsed I–V simulations on two different quiescent bias conditions at constant ambient temperature.
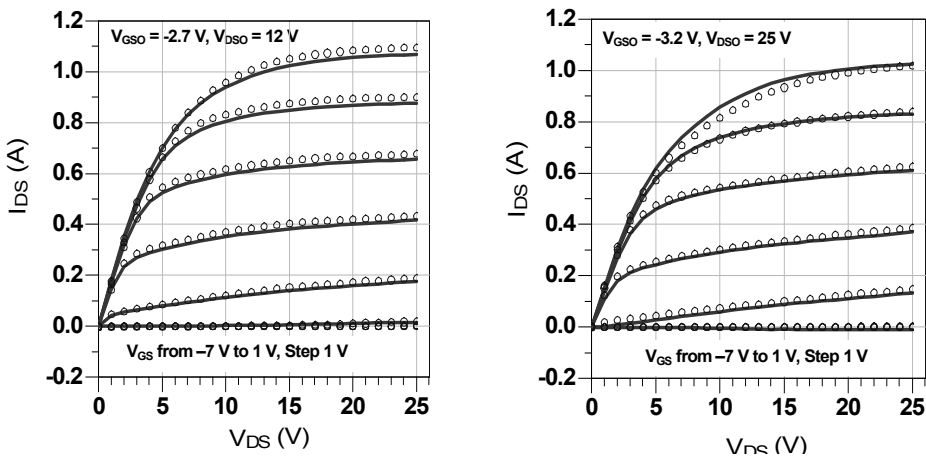
Fig. 18 (Lines) Pulsed I–V simulations and (circles) measurements for a 8x125-µm GaN HEMT at different quiescent bias conditions. © 2006 IEEE. Reprinted with permission.
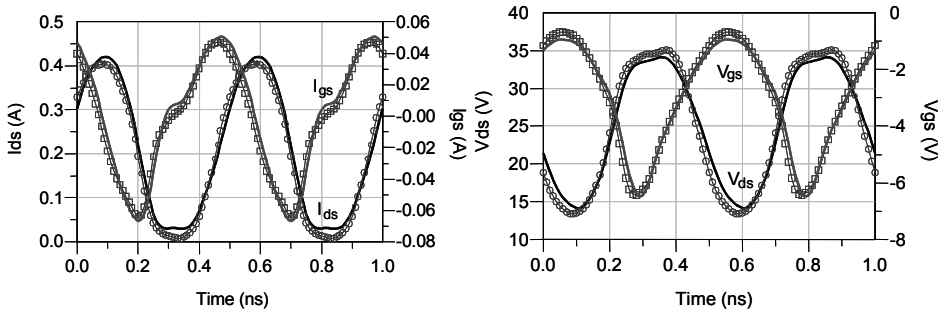
Fig. 19 (Lines) Simulated and (symbols) measured large-signal waveforms for class-AB-operated 8x125-µm GaN HEMT at 16-dBm input power. © 2006 IEEE. Reprinted with permission.

The very good agreement between simulation and measurement shows the ability of the model for describing the bias dependence of the trapping and self-heating effects. In

addition, these simulations verify the convergence behaviour of the model response under pulsed stimulation, which is very important for digital applications. Large-signal waveform measurements for 8×125-µm GaN HEMTs were done using the measurement setup described in (Raay & Kompa, 1997) and then simulated by the model. As it can be seen in Figure 19, very good agreement between measured and simulated current and voltage waveforms is obtained. This can be related to the improved construction of the model elements using the spline-approximation technique, as explained in Section 4.1, which improves the modeling of the higher order harmonics.
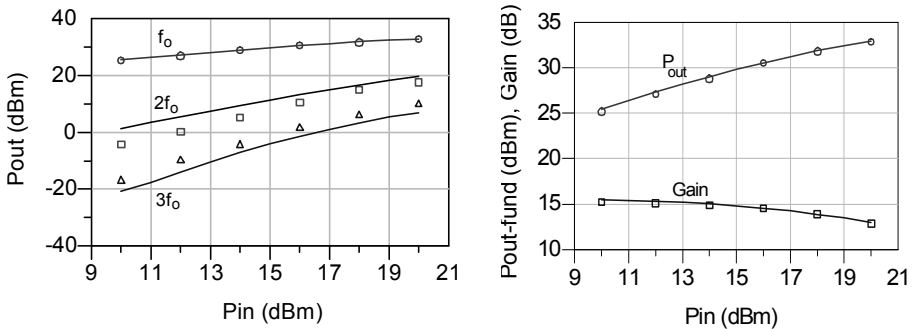
Fig. 20. (Lines) Single-tone power-sweep simulations compared with (symbols) measurements for class-A-operated 8x125-µm GaN HEMT at 2 GHz in a 50-Ω source and load environment. © 2007 IEEE. Reprinted with permission.

Figure 20 shows a simulation result of a single-tone input-power sweep for a 8×125-µm GaN HEMT. The model shows very good results with respect to the fundamental output power and gain even for input-power levels beyond the 1-dB gain-compression point. The model also shows good simulation results for the output power of higher harmonic components up to the third harmonic.
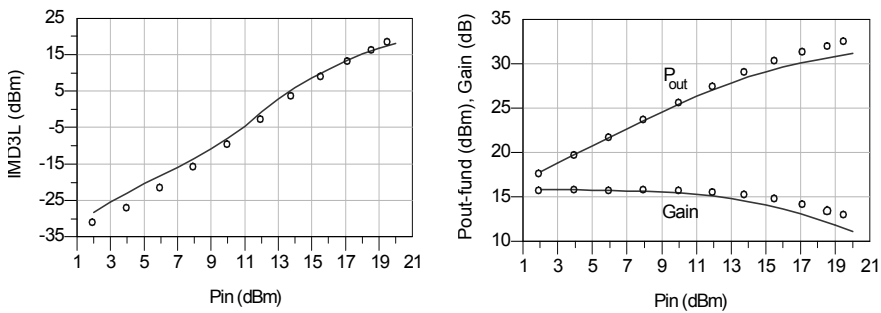
Fig. 21. (Lines) Simulated and (symbols) measured Pout, Gain and IMD3 versus input power per tone under two-tone excitation centered at 2 GHz and separated by 100 kHz for class-AB-operated 8x125-µm GaN HEMT in a 50-Ω source and load environment. © 2007 IEEE. Reprinted with permission.

Simulations for output power, gain and third intermodulation distortion under two-tone excitation centered at 2 GHz and separated by 100 kHz were performed. The simulation results are compared with measurements of 8×125-µm GaN HEMTs on different wafers. These measurements were performed using the developed measurement setups described in (Ahmad et al., 2005). Figure 21 presents the simulation results in comparison with the measurements. The model shows very good results for describing the output power and gain except at high-power end. The inaccuracy is due to the extrapolation error outside the region of measurements where the model was derived from. The model accuracy can be improved by increasing the range of these measurements to cover higher voltage conditions. The model also shows very good simulation for the third-order IMD. This can also be related to the use of spline approximation for the construction of the model-element data.
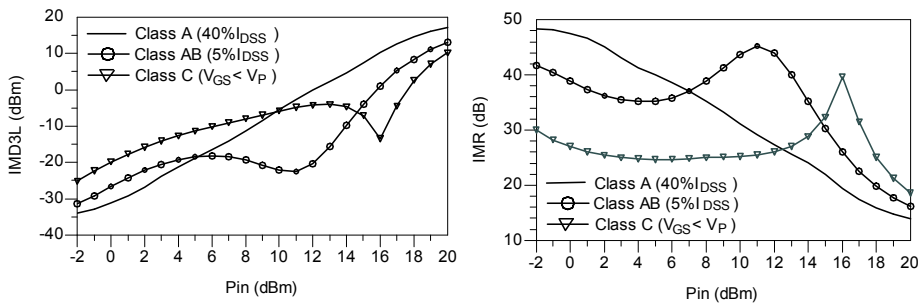
Fig. 22 Simulated lower intermodulation distortion and carrier to intermodulation ratio versus input power per tone under two-tone excitation centered at 2 GHz and separated by 100 kHz for a 8x125-µm GaN HEMT under 20 V drain bias voltage for different gate bias voltages in a 50 Ω source and load environment.

Figure 22 shows simulated lower IMD3 and the corresponding carrier to intermodulation ratio (IMR) for 8x125 µm GaN HEMT under two-tone excitation for different classes of operation. The model shows very good results for prediction of the IMD3 sweet spots (local minima), which result from the interaction between small- and large signal IMDs (Carvalho & Pedro, 1999; Fager et al., 2002). The IMD3 simulation is done at different gate bias conditions for 20 V drain biased device in a 50-Ω source and load environment. It is found that the best performance with maximum IMR and high power added efficiency could be obtained when the device is biased just above the pinch-off voltage as illustrated in Figure 22. These results are in a very good agreement with the reported ones in (Cabral et al., 2004) for a 2-mm gate width GaN HEMT.

5. Conclusion

In this chapter, a large-signal model for GaN HEMTs, which accurately predicts trapping- and self-heating-induced current dispersion and IMD, was developed and demonstrated. Detailed procedures for both small-signal and large-signal model parameter extraction has been presented. The extracted intrinsic gate capacitances and conductances of distributed small-signal model were integrated to find the gate charge and current sources of the large-signal model, assuming that these elements satisfy the integral path-independence

condition. Pulsed I–V measurements under appropriate quiescent bias conditions were used to accurately characterize and model the drain current and the inherent self-heating and trapping effects. It is found that using approximation technique for the construction of the large-signal-model database can improve the model capability for harmonics and IMD simulations. Large-signal simulations show that the model can accurately describe the performance of the device under constant external temperature. However, this model can also be extended to consider the variation of the ambient temperature.

6. References

Ambacher, O.; Smart, J.; Shealy, J.; Weimann, N.; Chu, K.; Murphy, M.; Schaff, W.; Eastman, L.; Dimitrov, R.; Wittmer, L.; Stutzman, M.; Rieger, W. and Hilsenbeck, J. (1999). Two-dimensional electron gases induced by spontaneous and piezoelectric polarization in undoped and doped AlGaN/GaN heterostructures. *Journal of Applied Physics*, Vol. 85, (March 1999) page numbers (3222-3232), ISSN 0021-8979.

Ahmed, A.; Srinidhi, E. & Kompa, G. (2005). Efficient PA modeling using neural network and measurement set-up for memory effect characterization in the power device, WE1D-5, ISBN 0-7803-8845-3, *Proceeding of International Microwave Symposium Digest*, USA, June 2005, Long Beach.

Cabral, P.; Pedro, J. & Carvalho, N. (2004). Nonlinear device model of microwave power GaN HEMTs for high power amplifier design. *IEEE Transaction Microwave Theory and Techniques*, Vol. 52, (November 2004) page numbers (2585-2592), ISSN 0018-9480.

Cuoco, V.; Van den Heijden, M. & De Vreede, L. (2002). The 'smoothie' data base model for the correct modeling of non-linear distortion in FET devices, *Proceeding of International Microwave Symposium Digest*, pp. 2149–2152, ISBN 0-7803-7239-5, USA, February 2002, IEEE, Seattle.

Charbonniaud, C.; De Meyer, S.; Quere, R. & Teyssier, J. (2003). Electrothermal and trapping effects characterization of AlGaN/GaN HEMTs, *Proceeding of European Gallium Arsenide & related III-V Compounds Application Symposium*, pp. 201-204, ISBN 1-58053-837-1, Germany, October 2003, Munich.

Carvalho, N. & Pedro, J. (1999). Large-and small-signal IMD behavior of microwave power amplifier. *IEEE Transaction Microwave Theory and Techniques*, Vol. 47, (December 1999) page numbers (2364-2374), ISSN 0018-9480.

Edwards, M. and Sinsky, J. (1992). A new criterion for linear two-port stability using a single geometrically derived parameter. *IEEE Transaction Microwave Theory and Techniques*, Vol. 40, (December 1992) page number (2303–2311), ISSN 0018-9480.

Eastman, L.; Tilak, V.; Smart, J.; Green, B.; Chumbes, E.; Dimitrov, R.; Hyungtak, K.; Ambacher, O; Weimann, N; Prunty, T; Murphy, M.; Schaff, W. & Shealy, J. (2001). Undoped GaN HEMTs for microwave power amplification. *IEEE Transaction on Electron Devices,* Vol. 48, (March 2001) page numbers (479-485), ISSN 0018-9383.

Filicori, F.; Vannini, G.; Santarelli, A.; Mediavilla, A.; Tazón, A. & Newport, Y. (1995). Empirical modeling of low-frequency dispersive effects due to traps and thermal phenomena in III–V FETs. *IEEE Transaction Microwave Theory and Techniques*, Vol. 43, (December 1995) page numbers (2972–2981), ISSN 0018-9480.

Fager, C.; Pedro, J.; Carvalho, N. & Zirath, H. (2002). Prediction of IMD in LDMOS transistor amplifiers using a new large-signal model, *IEEE Transaction Microwave Theory and Techniques*, Vol. 50, (December 2002) page numbers (2834-2842), ISSN 0018-9480.

Green, B.; Chu, K.; Kim, H.; Lin, H.; Tilak, V.; Shealy, J.; Smart, A. & Eastman, L. (2000). Validation of an analytical large signal model for GaN HEMTs, *Proceeding of International Microwave Symposium Digest*, pp. 761-764, ISBN 0-7803-5687-X, USA, June 2000, IEEE, Boston.

Goyal, R. (1989). *Monolithic Microwave Integrated Circuit: Technology and Design*, Artech House, ISBN 0890063095, Norwood, USA.

Hansen, P.; Strausser, Y.; Erickson, A.; Tarsa, E.; Kozodoy, P.; Brazel, E.; Ibbetson, J.; Mishra, U.; Narayanamurti, V.; DenBaars, S.; and Speck, J. (1998). Scanning capacitance microscopy imaging of threading dislocations in GaN films grown on (0001) sapphire by metalorganic chemical vapor deposition. *Applied Physics Letters*, Vol. 72, (May 1998) page numbers (2247–2249), ISSN 0003-6951.

Jarndal, A. & Kompa, G. (2005). A new small-signal modeling approach applied to GaN devices. *IEEE Transaction Microwave Theory and Techniques*, Vol. 53, (November 2005), page numbers (3440–3448), ISSN 0018-9480.

Jarndal[a], A. & Kompa, G. (2006). An accurate small-signal model for AlGaN-GaN HEMT suitable for scalable large-signal model construction. *IEEE Microwave and Wireless Components Letters*, Vol. 16, (June 2006) page numbers (333–335), ISSN 1531-1309.

Jarndal[b], A.; Bunz, B. & Kompa, G. (2006). Accurate large-signal modeling of AlGaN-GaN HEMT including trapping and self-heating induced dispersion, *Proceeding IEEE International Symposium Power Semiconductor Devices and ICs*, pp. 1–4, ISBN 0-7803-9714-2, Italy, June 2006, Napoli.

Jarndal, A. & Kompa, G. (2007). Large-Signal Model for AlGaN/GaN HEMT Accurately Predicts Trapping and Self-Heating Induced Dispersion and Intermodulation Distortion. *IEEE Transaction Microwave Theory and Techniques*, Vol. 54, (November 2007) page numbers (2830-2836), ISSN 0018-9480.

Kohn, E.; Daumiller, I.; Kunze, M.; Neuburger, M.; Seyboth, M.; Jenkins, T.; Sewell, J.; Norstand, J.; Smorchkova, Y. & Mishra, U. (2003). Transient characteristics of GaN-based heterostructure field-effect transistors. *IEEE Transaction Microwave Theory and Techniques*, Vol. 51, (Februry 2003) page numbers (634–642), ISSN 0018-9480.

Koh, K.; Park, H.-M. & Hong, S. (2002). A spline large-signal FET model based on bias-dependent pulsed I–V measurements. *IEEE Transaction Microwave Theory and Techniques*, Vol. 50 (November 2002) page numbers (2598–2603), ISSN 0018-9480.

Kotzebue, K. (1976). Maximally Efficient Gain: A Figure of Merit for Linear Active 2-Ports. *Electronics Letters*, Vol. 12, (September 1976) page numbers (490-491), ISSN 0013-5194.

Kompa, G. & Novotny, M. (1997). Frequency-dependent measurement error analysis and refined FET model parameter extraction including bias-dependent series resistors, *Proceeding of International IEEE Workshop on Experimentally Based FET Device Modelling and Related Nonlinear Circuit Design*, pp. 6.1–6.16, Report Number: A864133, University of Kassel, Germany, July 1997, IEEE, Kassel.

Lossy[a], R.; Chaturvedi, N.; Heymann, P.; Würfl, J.; Müller, S.; and Köhler, K. (2002). Large area AlGaN/GaN HEMTs grown on insulating silicon carbide substrates. *Physica Status Solidi (a)*, Vol. 194, (December 2002) page numbers (460–463), ISSN 0031-8965.

Lossy, R.; Hilsenbeck, J.; Würfl, J. and Obloh, H. (2001). Uniformity and scalability of AlGaN/GaN HEMTs using stepper lithography. *Physica Status Solidi (a)*, Vol. 188, (November 2001) page numbers (263–266),), ISSN 0031-8965.

Lossy[b], R.; Heymann, P.; Würfl, J.; Chaturvedi, N.; Müller, S.; and Köhler, K. (2002). Power RF-operation of AlGaN/GaN HEMTs grown on insulating silicon carbide substrates, *Proceeding of European Gallium Arsenide & related III-V Compounds Application Symposium*, ISBN 0-86213-213-4, Italy, September 2002, IEEE, Milan.

Lee, J. & Webb, K. (2004). A temperature-dependent nonlinear analytic model for AlGaN-GaN HEMTs on SiC. *IEEE Transaction Microwave Theory and Techniques*, Vol. 52, (January 2004) page numbers (2-9), ISSN 0018-9480.

Lin, L. & Kompa, G. (1994). FET model parameter extraction based on optimization with multiplane data-fitting and bidirectional search—A new concept. *IEEE Transaction Microwave Theory and Techniques*, Vol. 42, (July 1994) page numbers (1114–1121), ISSN 0018-9480.

Lu, S.-S.; Chen, T.-W.; Chen, H.-C.; Meng, C. (2001). The origin of the kink phenomenon of transistor scattering parameter S_{22}. *IEEE Transaction Microwave Theory and Techniques*, Vol. 49, (February 2001) page numbers (333 - 340), ISSN 0018-9480.

Meneghesso, G.; Verzellesi, G. ; Pierobon, R. ; Rampazzo, F.; Chini, A.; Mishra, U.; Canali, C. & Zanoni, E. (2004). Surface-related drain current dispersion effects in AlGaN-GaN HEMTs. *IEEE Transaction Microwave Theory and Techniques*, Vol. 51, (October 2004) page numbers (1554-1561), ISSN 0018-9480.

Root, D.; Fan, S. & Meyer, J. (1991). Technology Independent Large Signal Non Quasi-Static FET Models by Direct Construction from Automatically Characterized Device Data, *Proceeding of European Microwave Conference*, pp. 927 – 932, EUMA.1991.336465, Germany, October 1991, IEEE, Stuttgart.

Raay, F.; Quay, R.; Kiefer, R.; Schlechtweg, M. & Weimann, G. (2003). Large signal modeling of GaN HEMTs with Psat > 4 W/mm at 30 GHz suitable for broadband power applications, *Proceeding of International Microwave Symposium Digest*, USA, PA, pp. 451-454, ISBN 0-7803-7695-1, June 2003, IEEE, Philadelphia.

Schmale, I. & Kompa, G. (1997). A physics-based non-linear FET model including dispersion and high gate-forward currents, *Proceeding of International Workshop on Experimentally Based FET Device Modelling and Related Nonlinear Circuit Design*, pp. 27.1-27.7, Report Number: A864133, University of Kassel, Germany, July 1997, IEEE, Kassel.

Schmale, I. & Kompa, G. (1997). An improved physics-based nonquasistatic FET-model, *Proceeding of European Microwave Conference*, pp. 328–330, ISBN 0-7803-4202-X, Jerusalem, September 1997, IEEE, Israel.

System Manual HP8510B Network Analyzer (1987). HP Company, P/N 08510-90074, USA, Santa Rosa, July 1987.

Vetury, R; Zhang, N; Keller, S & Mishra, U. (2001). The impact of surface states on the DC and RF characteristics of GaN HFETs. *IEEE Transaction on Electron Devices*, Vol. 48, (March 2001) page numbers (560-566), ISSN 0018-9383.

Van Raay, F. & Kompa, G. (1997). Combination of waveform and load-pull measurements, *Proceeding of International IEEE Workshop on Experimentally Based FET Device Modelling and Related Nonlinear Circuit Design*, pp. 10.1-10.11, Report Number: A864133, University of Kassel, Germany, July 1997, IEEE, Kassel.

Wei, C.; Tkachenko, Y. & Bartle, D. (1999). Table-based dynamic FET model assembled from
      small-signal models. *IEEE Transaction Microwave Theory and Techniques*, Vol. 47,
      (June 1999) page numbers (700–705), ISSN 0018-9480.

Polyphase Filter Design Methodology for Wireless communication Applications

Fayrouz Haddad, Lakhdar Zaïd, Wenceslass Rahajandraibe
and Oussama Frioui
IM2NP – University of Provence
Marseille - France

1. Introduction

The growing wireless communication market has generated increasing interest in highly integrated circuits. This has been a relentless pressure for low cost, low power and small size of transceivers. At the same time, the emergent mobile communications require high-speed data transmission and high data-rate systems. For instance, the IEEE 802.11a/g wireless standards, which incorporate OFDM (Orthogonal Frequency Division Multiplexing) modulation, are able to provide up to 54Mbps. This has led to many improvements in design and development of circuit components and transceiver architectures.

Progress in silicon integrated circuit technology and innovations in their design have enabled mobile products and services. In most current designs, the analog part of a receiver uses multiple integrated circuits which may have been implemented in Gallium Arsenide (GaAs), Silicon Germanium (SiGe) or bipolar processes. They offer the best performance in terms of speed, noise sensitivity and component matching. Implementing similar circuits in CMOS processes with the same performances is still challenging because of its low process tolerances, parasitic effects and low quality factor passive components. However, CMOS process offers high density of integration and low consumption making it a good candidate for wireless communications (Mikkelsen, 1998).

In radiofrequency (RF) receivers, frequency down-conversion is an essential operation. It consists on the translation of the incoming RF signal to a lower frequency called the intermediate frequency (IF). This is typically performed by mixing the amplified RF signal with the local oscillator (LO) signal. The IF is defined as

$$f_{IF} = |f_{RF} - f_{LO}| \qquad (1)$$

However, this frequency translation provides a serious problem of frequency image rejection (Razavi (a), 1997). Hence, classical wireless receiver architectures have been commonly implemented using the superheterodyne topology, in which the image suppression is done by off-chip devices such as discrete components, ceramic or surface acoustic wave (SAW) filters (Razavi, 1996; Samavati et al, 2000; Macedo & Copeland, 1998). They have high quality factor and good linearity; however, their high cost and their non

integration make them less attractive to be used in the emerging integrated receivers (Razavi (a), 1997; Huang et al, 1999).

To overcome this drawback, zero-IF receiver architectures, in which the RF signal is transposed directly to baseband, have been proposed (Razavi (b), 1997; Behzad et al, 2003). Since the LO is at the same frequency as the RF input, this architecture removes the IF and the image rejection problem, which arises differently in the receiver chain and results from mismatches between the I (in-phase) and Q (quadrature) paths as well as amplitude mismatches. Although the direct conversion performs well image rejection, this architecture suffers from flicker noise, DC offsets and self-mixing at the inputs of the mixers, resulting in filter saturation and distortion.

To understand how the problem of frequency image arises, consider the process of down-conversion as represented in the Fig.1. When mixing the wanted signal band (at f_{RF}) with the LO, the obtained signal band is located at f_{IF}. But, since a simple analog multiplication does not preserve the polarity of the difference between the two mixed signals (i.e $\cos(\omega_1-\omega_2)t=\cos(\omega_2-\omega_1)t$), the signal band at $f_{RF}-2f_{IF}$ is also translated to the same IF after mixing with the LO. The signal at $f_{RF}-2f_{IF}$ is known as the image frequency. Therefore, any undesired signal located at the image frequency will be translated to the same IF along with the desired RF signal. And, this image signal may distort the wanted signal and lead to an improper system work. Thus, the image signal must be filtered before mixing.
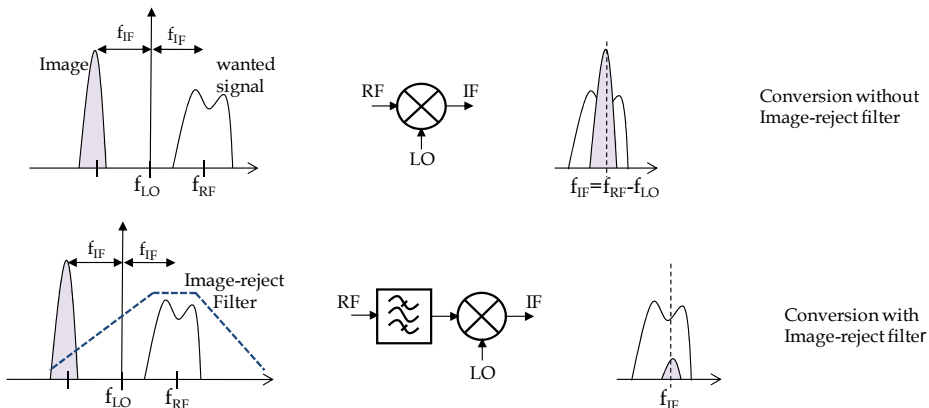
Fig. 1. Image rejection problem in RF receivers

An important specification to determine the performance of a receiver and to quantify its degree of image rejection is the image rejection ratio defined as the ratio of the magnitude in the attenuation band to that in the passband and can be given by

$$IRR = \frac{Desired\ Signal\ Level}{Image\ Signal\ Level} \tag{2}$$

The IRR required to ensure signal integrity and suitable bit-error-rate (BER) varies depending on the application. As an example, for short-range applications where low or moderate selectivity is required, an image suppression of 45dB is adequate, but is far less than that required in long-range heterodyne receivers. For example, DECT and DCS-1800

applications require 50dB and 60dB of image rejection respectively (Long, 1996). Method proposed ten years ago (Rudell et al, 1997), enforced external tuning or laser trimming and achieved image-rejection ratios, typically on the order of 35-50dB.

In this article, we will study first a state-of-the-art of the image rejection techniques as well as their implementation constraints inside wireless architectures. This study concerns essentially the image-reject architectures and focuses on complex polyphase filters. Then, we will propose a design methodology dedicated to passive polyphase filters (PPF) which includes in the design flow an analytical model allowing quantifying the impact of the mismatch of the components and the resulting signal in the IRR degradation. This methodology takes into account the non ideality of passive components (parasitic capacitances and resistors) together with the symmetry of the signal path during the layout design. Different techniques dedicated to layout matching combined with optimum component sizing from an experimental method are proposed so as to increase the IRR. Such a method gives a possibility to design PPFs operating from wide frequency range (1MHz to 5GHz) and allows attaining high performances in terms of IRR (about 60 dB). The proposed method has been validated with some test-cases in full CMOS technology.

2. Image rejection techniques

2.1 Image-reject architectures

Image-rejection architectures are the most known methods for implementing image rejection structures. They can typically be divided into half-complex and full-complex architectures (Crols et al, 1998; Steyaert et al, 2000) (Fig.2).
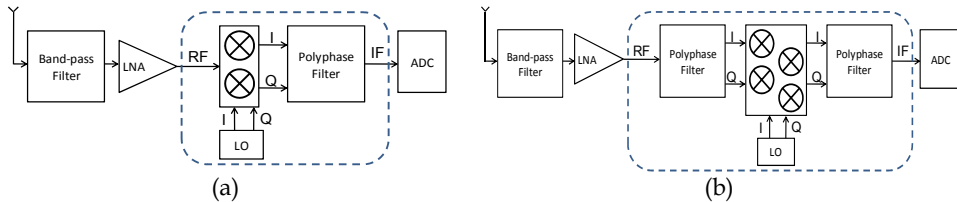
Fig. 2. Half-complex (a) and full-complex (b) receiver architectures, using polyphase filters

The half-complex architecture is based on the use of image-reject mixers combined with passive or active filters. As shown in Fig.2(a), a real RF signal is mixed with a LO complex signal to feed the IF polyphase filter. The quality of the image rejection inside such an architecture results mainly from three parameters: i) the balance between I and Q signals (phase and magnitude error), ii) the adequate matching of mixers iii) the polyphase filter performances.

There are two well-known architectures using such techniques: the Hartley and Weaver architectures, depicted in Fig.3 (Razavi (b), 1996; Xu et al, 2001). Generally, the Weaver topology is preferred to the Hartley architecture. In fact, the 90° phase shifter bloc (Fig.3(a)) comes with hard design constraints in terms of component matching which result to significant phase error especially at high frequencies. For instance, a change of 20% in resistors and capacitors (used to generate the quadrature), due to temperature and process variations, gives an IRR of only 20dB (Maligeorgos & Long, 2000).
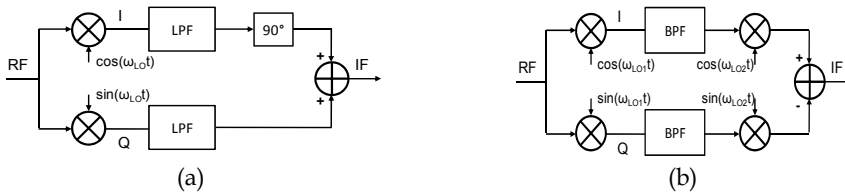
(a)   (b)

Fig. 3. Hartley (a) and Weaver (b) image-reject architectures

Basic Hartley and Weaver implementations proposed in the literature have typically IRRs in the range of 30-40dB (Carta et al, 2005). This is far below the 60dB required by wireless standards (Long & Maliepaard, 1999). In fact, phase mismatch between I and Q signals and gain mismatch between mixer signal paths result to much lower image rejection value. The IRR can be expressed as a function of the mismatches (Rudell et al, 1997)

$$IRR = \frac{1 + (1 + \Delta A)^2 + 2(1 + \Delta A)\cos(\varphi_1 + \varphi_2)}{1 + (1 + \Delta A)^2 - 2(1 + \Delta A)\cos(\varphi_1 + \varphi_2)} \tag{3}$$

where φ_1 and φ_2 represent the phase errors of LO_1 and LO_2 respectively, and ΔA is the gain error between I and Q paths. As shown in Fig.4, the IRR as a function of the total phase and gain mismatch. Thus, in order to provide an IRR of 60dB with a gain mismatch of 0.1%, the LO phase errors must be less than 0.1°.
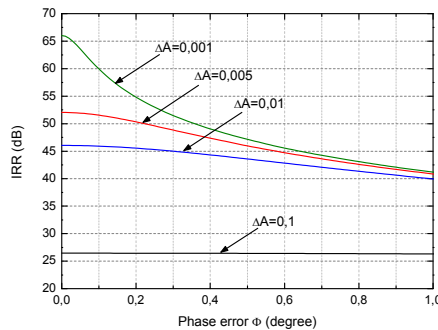
Fig. 4. Image rejection ratio (IRR) versus the phase mismatch for different gain mismatch quantities (Rudell et al., 1997)

The need for a monolithic solution for image-reject receiver which can perform gain and phase calibration is obvious. Various calibration techniques to correct the mismatch have been developed using analog or digital circuits. Digital calibration techniques have been implemented at the analog-to-digital converter (ADC) level inside the receiver chain (Valkama & Renfors, 2000; Sun et al, 2008) using either background digital correlation loops or commutated feedback capacitor switching to correct non-idealities in baseband components. Another technique is proposed in (Montemayor & Razavi, 2000) and consists on a self-calibrating architecture. It determines the phase and gain mismatches of a Weaver architecture and applies a feedback through gain and phase adjustment. A special dedicated tone at the image frequency is used and a periodic calibration with the external image tone is needed. This property restricts its application on systems using time division multiple access (TDMA). This method improves the image rejection with no penalty in linearity,

noise or gain, but increases the power consumption. Another technique consists on continuous calibration in a modified-image-reject-Weaver architecture (Elmala & Embabi, 2004). The phase and gain mismatches quantities are calibrated independently without the use of any external calibrating tones. This system generates two control signals using a variable delay gain circuit in two independent calibration loops. A further technique, based on the Weaver architecture, consists on simultaneous calibration by a sign-sign least mean square (SS-LMS) algorithm (Der & Razavi, 2003). The LMS adaptation circuit adjusts the phase and gain mismatches differentially to avoid systematic control. Digitally storing the calibration coefficients solves the problem of periodic refreshing, making it suitable for systems using code division multiple access (CDMA), but an external image tone is still needed in calibration procedure. Furthermore, a post-processing image rejection algorithm is proposed in (Lerstaveesin & Song, 2006) to reject the image in the baseband using an adaptive zero-forcing sign-sign feedback concept and does not require complicated digital processing. This algorithm can detect and correct I/Q imbalance continuously, but it alleviates the need for a high resolution ADC in the digital image rejection device.

Despite the difficulty to realize accurate phase shifters, (Chou & Lee, 2007) demonstrates that this is not essential in the Weaver architecture. The image rejection is performed by making the phase mismatch between I and Q signals of the first LO to be equal to that of the second LO. Thus the design constraints on phase matching are relaxed, and more attention can be placed on gain matching (Chou & Lee, 2007).

The full-complex architecture (also referred as double quadrature down-conversion) requires the use of complex polyphase filters. The complex polyphase filters are suitable for high frequency applications since they can meet the dynamic range and bandwidth requirement in RF frequencies (Wu & Chou, 2004). In this case, a notch frequency located at the image frequency is used to reject image signals rather than bandpass filtering. As shown in Fig.2(b), the RF signal is complex filtered (RF polyphase filter), then the RF and LO complex signals are multiplied together to feed the IF polyphase filter, to reach about 60dB of image suppression (Behbahani et al, 2001). The interest of this structure comes from the fact that the image rejection is supported in the RF domain by the RF polyphase filter and the quadrature LO, which is advantageous compared to the half-complex architecture. Thus, the design constraints in terms of image rejection are relaxed in the RF polyphase filter and the LO compared to the IF polyphase filter.

Summary of performances of numerous image-reject architectures reported above is given in table 1.

	Technology	RF/IF (Hz)	IRR	IIP3	NF	Power consumption
(Rudell et al., 1997)	0.6μm CMOS	1.9G / 200M	45dB	-7 dBm	14dB	92mW
(Carta et al., 2005)	BiCMOS	5 - 2.4 G/ 20M	33dB	-12 dBm	8.9dB	19mW
(Behbahani et al., 2001)	0.6μm CMOS	270M / 10M	60dB	-	-	62.7mW

	Technology	RF/IF (Hz)	IRR	IIP3	NF	Power consumption
(Crols & Steyaert, 1995)	0.7µm CMOS	900M /3M	46dB	27.9 dBm	24dB	500mW
(Wu & Razavi, 1998)	0.6µm CMOS	900M / 400M	40dB	-8 dBm	4.7dB	72mW
(Banu et al., 1997)	0.5µm BiCMOS	900M / 10.7M	50dB	-4.5 dBm	4.8dB	60mW
(Lee et al., 1998)	0.8µm CMOS	1G/ 100M	29dB	0.6 dBm	19dB	108mW
(Behbahani et al., 1999)	0.6µm CMOS	270M/ 10M	58.5 dB	-8 dBm	6.1dB	33mW
(Samavati et al., 2001)	0.24µm CMOS	5.2G	53dB	-7dBm	7.3dB	58.8mW
(Meng et al., 2005)	GaInP/ GaAs HBT	5.2G / 30M	40dB	-10dBm	-	150mW
(Wu & Chou, 2003)	0.18µm CMOS	5G / 20M	50.6dB	-13dBm	8.5dB	22.4mW
(Kim & Lee, 2006)	0.18µm CMOS	5.25G/1G	40dB	-8dBm	7.9dB	57.6mW
(Razavi, 2001)	0.25µm CMOS	5.25G/2.6G	62dB	-15dBm	6.4dB	29mW
(Lee et al., 2002)	0.25µm CMOS	5.25G/ 300M	51dB	-7dBm	7.2dB	21.6mW
(Chou & Wu, 2005)	0.25µm CMOS	6M - 30M	48dB	-8dBm	-	11mW

Table 1. Circuit performances using the Weaver and double quadrature conversion architectures

2.2 Complex polyphase filters

A Hilbert filter responds to the complex representation of a signal and is based on a shift transform, $s \mapsto (s + j\omega_0)$ (Khvedelidze, 2001). It translates the poles and transforms the lowpass response into a bandpass response centered at $\omega=\omega_0$, while preserving both amplitude and phase characteristics. Thus, owing to its asymmetric response to positive and negative frequencies, such a filter may be synthesized to suppress the image and pass the desired frequency; as the case of polyphase filters (Chou & Wu, 2005).

Invented by Gingell in 1971, polyphase filters were used to generate of quadrature signals in audio applications and were implemented first using discrete components (Gingell, 1971). This work has many limitations since working on such low frequency audio domain does not consider the influences of parasitic resistors and capacitors, moreover, components mismatch was not analyzed in the discrete components implementation (Tetsuo, 1995). Integrated PPFs were rediscovered in 1994 as an efficient RF quadrature generation technique in CMOS technology (Steyaert & Crols, 1994). The design of integrated CMOS PPF faces many challenges, so that many researches aim to analyze the sensitivity of the RF CMOS PPF in RF integrated transceivers (Galal & Tawfik, 1999) and their application in image rejection. This analysis allows understanding the PPF behavior, but it remains too theoretical for designers to get quantitative results about influences of process and mismatch variations on PPF performances.

A polyphase signal is a set of two or more vectors having the same frequency but different in phase (Galal et al, 2000). If its vectors have the same magnitude and are equally spaced in phase, it is considered symmetric. Hence, a symmetric two-phase signal consists of two vectors of equal magnitudes with the same frequency and being separated in phase by 180°. The phase order of the signal vectors determines the polarity of the polyphase signal sequence, i.e. a positive sequence has a clockwise phase order, while a negative sequence has an anticlockwise phase order. This introduces the concept of negative and positive frequencies (Fig.5). It should be noted that the phase order is different from the direction of rotation because all sequences, whether positive or negative, consist of vectors rotating anticlockwise. Since PPF networks have asymmetric responses to inputs of opposite polarities, they were described as *asymmetric* (Tetsuo, 1995).

The study of the PPF response can be performed by the way of vector analysis (Galal & Tawfik, 1999). Since the PPF phases are symmetric, the chain matrix of a single phase represents the chain matrix of the network. The PPF output is considered as the sum of the outputs cascaded by each symmetric input alone thanks to linear superposition rules. Fig.6 shows the structure of one phase generalized PPF, where admittance Y1 is connected between the input and the corresponding output, and Y2 is skewed between the input and output of adjacent phases. The chain matrix of a single phase can be written as (Galal & Tawfik, 1999)

$$\begin{bmatrix} Vin,k \\ Iin,k \end{bmatrix} = \frac{1}{Y1+e^{j\theta}Y2}\begin{bmatrix} Y1+Y2 & 1 \\ 2.Y1.Y2(1-cos\theta) & Y1+Y2 \end{bmatrix}\begin{bmatrix} Vout,k \\ Iout,k \end{bmatrix} \tag{4}$$

where θ represents the relative phase difference between V_{in} and the neighboring inputs, which in turn determines the polarity of the inputs. If $\theta<0$, V_{in} will be leading $e^{j\theta}V_{in}$ in phase, which causes the inputs to be positive. On the other hand, if $\theta>0$, V_{in} will be lagging, thus causing the inputs to be negative.
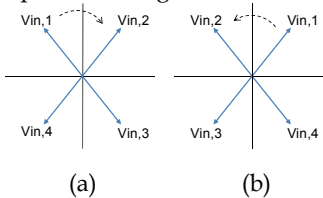

(a)	(b)

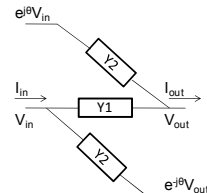Fig. 5. Two polyphase signals has a positive phase sequence (a) and a negative phase sequence (b)

Fig. 6. A single phase of a generalized PPF network

Resistance-capacitance (RC) polyphase networks are a special case of sequence asymmetric polyphase networks. They represent a passive implementation of polyphase filters which makes them attractive for integrated high-frequency and low-power applications. The structure of a four-phase RC polyphase filter network, which composes one stage, is shown in Fig.7. One stage of an RC PPF contributes to one pole frequency, called a *notch*, around which the image is attenuated. It can be derived as

$$f_p = \frac{1}{2\pi RC} \tag{5}$$
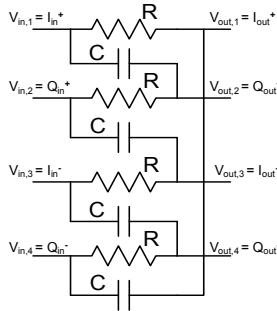
Fig. 7. One "four-phase" RC polyphase filter stage

3. Passive Polyphase Filter (PPF) Design

3.1 Case of a single-stage PPF

Let us now express the transfer function H(ω) of the PPF. Since the PPF is a complex filter, its transfer response can be represented as

$$H(\omega) = H_1(\omega) + j.H_2(\omega) \tag{6}$$

where $H_1(\omega)$ and $H_2(\omega)$ are the real and imaginary parts of the transfer function respectively. Let us note H(ω) the transfer function for any phase in a positive sequence, while H(-ω) that of negative one. Since complex polyphase filters discriminate positive and negative sequences, H(ω) ≠ H(-ω).

In the case of four-phase RC PPF shown in Fig.7, and substituting in (4) Y1, Y2 and θ for 1/R, sC and ±π/2 respectively, the open-circuit ($I_{out,k}$=0) voltage transfer function in a positive sequence can be written such as

$$H(\omega) = \frac{V_{out,k}}{V_{in,k}} = \frac{\omega_p - \omega}{\omega_p + j\omega} = A\frac{1 - \omega RC}{1 + j\omega RC} \tag{7}$$

while in a negative sequence, we yield

$$H(-\omega) = \frac{V_{out,k}}{V_{in,k}} = \frac{\omega_p + \omega}{\omega_p + j\omega} = A\frac{1 + \omega RC}{1 + j\omega RC} \tag{8}$$

where *A* is an amplification factor and ω_p is the pole frequency. The transfer curves of |H(f)|$_{dB}$ and |H(-f)|$_{dB}$ are shown in Fig.8 for f_p =2.4GHz. Note that the desired signal with positive frequency falls in the filter's passband while the image signal at negative frequency is attenuated. The IRR, previously defined in (2), can be expressed as following

$$IRR(\omega) = \frac{|H(\omega)|}{|H(-\omega)|} = \frac{\omega_p - \omega}{\omega_p + \omega} \tag{9}$$

Thus, according to (9), we can deduce that IRR (f_P) = 0 if $H_1(\omega)$ and $H_2(\omega)$ are perfectly matched. Nevertheless, this theoretical value is difficult to achieve due to intrinsic mismatch of the components and the non uniformity of the connection lines.

Now, let us generalize the transfer function considering gain and pole mismatches which can be expressed as

$$H_m(s) = \frac{A\omega_p \left(1 - \frac{\Delta A}{2A}\right)\left(1 - \frac{\Delta\omega_p}{2\omega_p}\right)}{s + \omega_p\left(1 - \frac{\Delta\omega_p}{2\omega_p}\right)} + j\frac{A\left(1 + \frac{\Delta A}{2A}\right)s}{s + \omega_p(1 + \frac{\Delta\omega_p}{2\omega_p})} \tag{10}$$

where ΔA and $\Delta\omega_p$ are the mismatch quantities of the gain and the pole frequency, respectively.

Analytical modeling has been performed in order to quantify the impact of the mismatch on the IRR degradation and the notch frequency drift. Results are summarized in table 2 for some typical values of ΔA and $\Delta\omega_p$. The same effect has been reported on Fig.8. It can be noted (table 2) that pole mismatch of 2% leads to an IRR degradation higher than 20dB and gain mismatch greater than 10% cause bandwidth degradation higher than 200MHz. It depicts that the gain mismatch ΔA shifts the frequency, while the pole frequency mismatch $\Delta\omega_p$ changes the IRR.
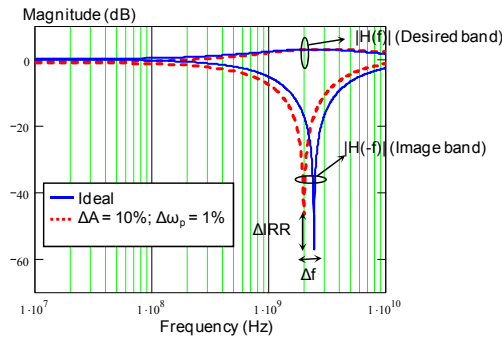
Fig. 8. Transfer responses of a PPF for a positive (|H(f)|) and negative (|H(-f)|) phase sequences

$\Delta\omega_p$(%)	ΔA (%)	Δf (MHz)	ΔIRR (dB)
1	20	500	10
1	10	200	10
1	5	50	10
2	5	50	20
3	5	50	23

Table 2. Gain and pole mismatch influence on the notch frequency and the IRR

This mismatch results essentially from components imbalance. Thus, considering a single RC PPF section, if the absolute values of R and C deviate such as R+ΔR and C+ΔC, the pole frequency differ from 1/(2πRC) and the residual voltage is given by

$$\delta_A = \frac{1 - \omega.\tau(1 + \delta_\tau)}{\sqrt{1 + (\omega.\tau(1 + \delta_\tau))^2}} \tag{11}$$

where δ_τ is the time constant fractional deviation (Behbahani et al, 2001).
Since in the vicinity of the pole frequency, $\omega.\tau = 1$, we have

$$\delta_A \simeq \frac{-\delta_\tau}{\sqrt{2}} \tag{12}$$

Further, the matching of a parameter M is generally defined as a standard deviation of Gaussian distribution of relative $\Delta M/M$ between identically designed paired devices. Then, the IRR with consideration of R and C mismatches can be expressed in a normalized RMS quantity (Behbahani et al, 2001) and then given by

$$\frac{\sigma(image\ Out)}{Desired\ Out} = \frac{\sigma_\tau}{4} = \frac{1}{4}\sqrt{\left(\frac{\sigma_R}{R}\right)^2 + \left(\frac{\sigma_C}{C}\right)^2} \tag{13}$$

3.2 Case of a multi-stage PPF

Since a one-stage PPF suppresses the image only around the notch frequency, it supplies a narrow band rejection. By cascading several stages with different notches, a wide bandwidth can be achieved. These notches have to be placed at equal frequency ratios (Fang et al, 2005); i.e

$$\alpha = \frac{f_{p2}}{f_{p1}} = \frac{f_{p3}}{f_{p2}}, etc. \tag{14}$$

Figure 9 depicts the topology of cascaded stages of RC polyphase filter and their correspondent frequency responses. It shows the image rejection through multi-stage-RC PPFs, in which the notches are logarithmically spaced for equiripple response. The bandwidth to be covered and the desirable image rejection amount fix the number of stages needed for the polyphase filter.
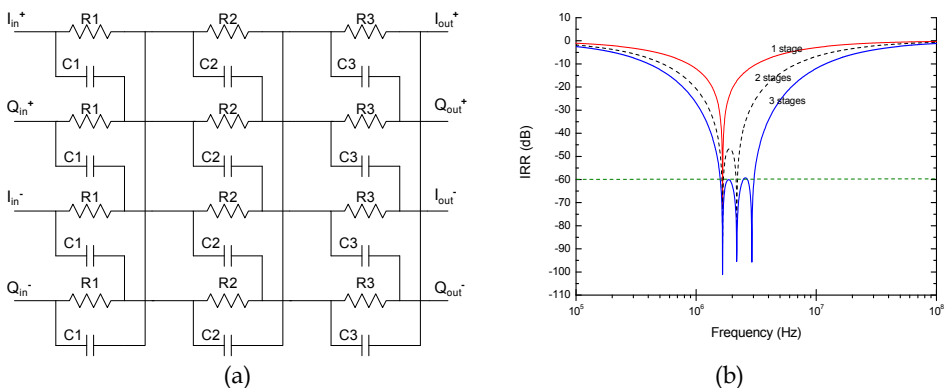
Fig. 9. A three-stage RC PPF topology (a); Cascade frequency responses of one-, two- and three-stage PPFs

In spite of its broadband, a multistage PPF presents gain losses because each stage loads the previous one. Thus, additional buffers should be included between the stages for loss compensation or impedance adaptation. The cost remains in the increased power consumption due to buffers. In addition, the pole location of the original RC PPF could be changed. Other structures of multistage RC polyphase filters are proposed in (Komoriyama et al, 2007) to overcome the problem of buffering and to ensure minimal element value-spread and equal-ripple gain.

In the case of a two-stage RC PPF, the transfer function can be obtained by multiplying the chain matrices, as given in (4), of each stage and evaluating the open-circuit ($I_{out,k}$=0) complex transfer function. We yield

$$H(\omega) = \frac{(1 - \omega R_1 C_1)(1 - \omega R_2 C_2)}{1 - \omega^2 (R_1 C_1 R_2 C_2) + j\omega (R_1 C_1 + R_2 C_2 + 2R_1 C_2)} \tag{15}$$

In the case of a three-stage RC PPF, the transfer function is given by

$$H(\omega) = \frac{(1 - \omega R_1 C_1)(1 - \omega R_2 C_2)(1 - \omega R_3 C_3)}{D(\omega)} \tag{16}$$

where $D(\omega) = 1 - \omega^2 (R_1 R_2 C_1 C_2 + R_1 R_3 C_1 C_3 + R_2 R_3 C_2 C_3 + 2R_1 R_3 C_2 C_3 + 2R_1 R_2 C_1 C_3 + 2R_1 R_2 C_2 C_3) + j\omega (R_1 C_1 + R_2 C_2 + R_3 C_3 + 2R_1 C_2 + 2R_1 C_3 + 2R_2 C_3) - j\omega^3 (R_1 R_2 R_3 C_1 C_2 C_3)$

The resultant IRR_N of N cascaded stages can be derived by multiplying the IRR of each one-stage polyphase filter as

$$IRR_N(\omega) = \prod_{n=1}^{N} IRR_n(\omega) \tag{17}$$

4. Active Polyphase Filters

In CMOS wireless receiver design, multi-stage polyphase filters are widely used. One solution consists on substituting the passive resistive parts of the filter by active one, using transconductances (Behbahani et al, 2002; Andreani et al, 2000). This topology behaves in the same way as the regular structure if gm equals to 1/R with ideal source and load. The high input impedance of the transconductance reduces the loading of the preceding stage. The active polyphase filter combines isolation, gain and small chip area.

Many realizations of active resistor are based on the use of inverter-type transconductor, originally proposed by *Nauta* (Andreani & Mattison, 2002). The transconductor circuit proposed in (Behbahani et al, 2000) has many advantages including good linearity, low-noise, high-frequency capability and low-voltage; but at the same time it suffers from common-mode instability. Further RC-active polyphase filter implementations are based on the use of second generation current conveyors (CCIIs) (Ün (a), 2004) or on the use of conventional operational amplifiers (OPAMPs) and RC components (Ün (b), 2004). However, in high-frequency operation, opamp RC filters design is problematic because of the required gain-bandwidth, the power consumption and the swing limitations. Thus, the OPAMP can be replaced by an operational transconductance amplifier (OTA) (Tsividis, 1994). Even so, problems due to OTAs design with adequate gain and bandwidth in low-voltage CMOS process are preserved. Hence, techniques of feedback and feedforward

common-mode compensation are used with such structures allowing high dc gain and good phase margin even in low-voltage CMOS applications (Harrison, 2002; Thandri & Silva-Martinez, 2003). Another structure of active RC PPF proposed in (Tillman & Sjoland, 2005) is based on CMOS inverters, with dc feedback to stabilize the bias point. It is used to generate quadrature signals and combines high gain and good quadrature performance (quadrature error<0.8° in the tuning range [9.14, 10.58] GHz).

Furthermore, (Chian et al, 2007) proposes a novel design idea to implement polyphase filters based on replacing passive components by MOSFETs. This active device gives the same functions as the conventional passive polyphase filter with a significant reduction of the chip area; but it includes great effects of nonlinearity and parasitic components, making it difficult to handle in the experimental plan. They can be realized also by using gyrators, but, it is difficult to realize a gyrator using practical passive elements because of its reciprocity. Other complex filters are reported as part of the receiver design and, therefore, details about the filter performance were not given (Van Zeijl et al, 2002).

The active polyphase filter solutions, comparing to the passive ones, have smaller area, making them more adequate for low and intermediate frequency applications, but have at the same time more power consumption and lower linearity. Owing to the recent improvements on CMOS technology, passive components present better quality, in particular in the high frequency domain. Then, it is more convenient to use PPFs in the RF part, with certainly a special attention to the parasitics and the matching. The electrical model used in EDA (CAO) tools is no more sufficient or not enough accurate to underline the parasitic contributions as well as mismatch effects while designing the RF PPFs. Therefore, it is necessary to perform a PPF modeling to achieve the suitable performances of the future wireless communication standards.

5. Mismatch analysis

While working with PPF, the image rejection depends on the ability of the designer to achieve sufficient matching on the resistors and capacitors which comes from many causes (Hastings, 2006) such as microscopic fluctuations in dimensions, process biases, diffusion interactions, thermoelectric effects, etc. The requirements from component matching are contradictory to that of minimization of noise coupling, signal loss and chip area. It is known from experimental studies (McNutt et al, 1994) that the variance of adjacent resistors and capacitors is inversely proportional to their area. Consequently, large component area is required to achieve high IRR, but, in the same time, large area increases the parasitic capacitances value of filter components. In fact, achieving high IRR with polyphase filter results from an optimal sizing of the filter components. In other terms, tradeoff between the chip area and IRR must to be considered. Furthermore, back-end design methodology including layout consideration is mandatory in order to optimize CMOS PPF.

Different simulations related to image rejection have been done to verify multi-stage RF PPFs for a given communication standard. In these experiments, let us consider image rejection in a low-IF receiver with RF PPFs working around 2.4GHz. First, the characteristics of the different stages and principally their notches frequencies are chosen. Once the notches frequencies are determined, values of resistors and capacitors can be selected. Small signal simulations with *SpectreRF* (*Cadence*®) have been considered to focus on the effect of the component variations on the IRR and consequently to calibrate judiciously the optimal sizes

and values of resistors composing the filter allowing the required IRR. To investigate this further, Fig.10 shows the simulated IRR results for different polyphase networks with mismatch consideration. The IRR is illustrated in three-dimensional plot as a function of the resistor's electrical value (R) and resistor size, which for the current study corresponds to the width (W). In the X-axis, the parameter R is used to calibrate the first stage of the polyphase filter. The resistor values of the other stages are set to a fixed pole ratio α, as shown in (14). The capacitors are chosen to give the right pole frequency.

These three-dimensional plots show first that multiplying the number of stages gives a higher IRR. For example for the couple (R, W) equal to (70Ω, 10μm), the IRR increases from 52dB with three-stage polyphase design (Fig.10(a)), to 60dB with four stages (Fig.10(b)) and reaches 65dB with five stages (Fig.10(c)). However, having many stages in the polyphase network conducts to a growth of the components number and increases the silicon area, the power loss and the parasitic capacitances. Hence, according to the costumer need, designers should make a compromise between achieving a polyphase filter with high image rejection and low area and low silicon area cost.

Furthermore, Fig.10 illustrates that a high IRR is achieved if the value and the size of the resistor converge to the optimal values on each multi-stage polyphase filter. For about the different filter configurations, it shows that the IRR variation versus R corresponding to a given configuration is quasi-linear. For instance for a five-stage PPF, the IRR changes from 65dB, to 68dB and 70dB for resistor's width of 10μm, 20μm and 40μm respectively (Fig.10(c)). In this case, it can be noted that a gain of only 5dB in the IRR produces an expansion of the resistor size by almost 400% confirming the existence of an optimal component sizing for a specified IRR with each polyphase filter configuration. The possible reason is that large component area yields better matching on the circuit and presents optimal parasitic capacitances effect.

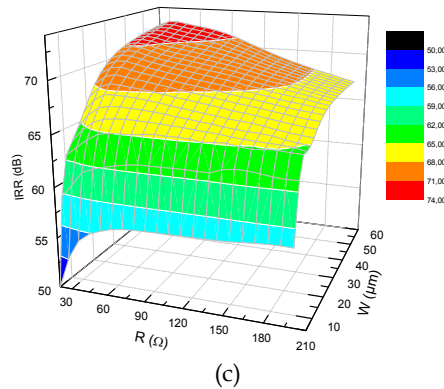(a)   (b)

(c)
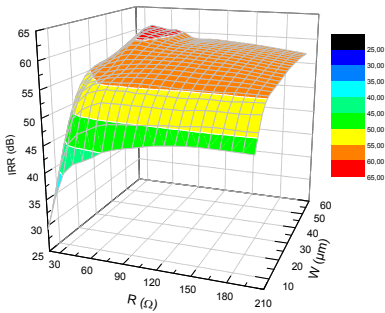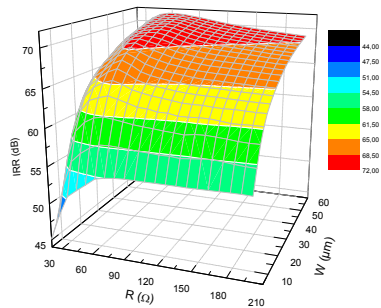
Fig. 10. IRR variation of (a) three, (b) four, and (c) five PPF versus resistor's sizes and values

A statistical representation is suitable to characterize the processes happening in probabilistic ways. In statistical simulations, sequences of random numbers with a certain probability distribution function are used to model the stochastic process. Usually, many statistical simulations runs are conducted and averaged to reach good accuracy of the simulation results. Process tolerances and component mismatch in integrated circuits are consequences of stochastic processes within a certain range, and they are usually available in CMOS process files derived by elaborate measurements. It is known that both process tolerances and component mismatch have truncated Gaussian probability distribution functions (Spence & Soin, 1997). In our application, Monte Carlo simulation can be applied to verify the statistical nature of the IRR with certain process tolerances and a resultant component mismatch, and to check the probability distribution of the gain mismatch. After optimal sizing and value calibrations of the PPF components as shown previously, three, four and five stages are simulated. The analysis concerns the process and mismatch variations of the PPF component corners (Polysilicon resistors and MIM (Metal-Insulator-Metal) capacitors for the current study) before parasitics extraction on the frequency band [2, 3] GHz. The Monte Carlo simulation results are expressed as frequency of occurrence histogram (5050 samples of RF PPFs) for different intervals of the IRR and shown in Fig. 11.

(a)   (b)   (c)
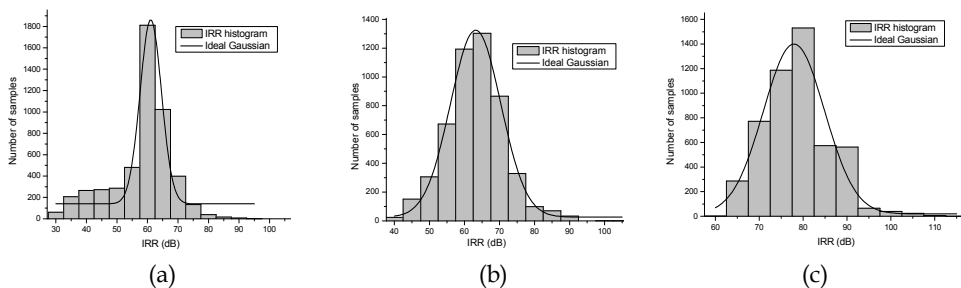
Fig. 11. Monte Carlo simulation results of (a) three, (b) four and (c) five stage RF PPF: IRR histogram (process and mismatch variations)

The impact of process and mismatch variations on the PPF response is summarized in table 3. It illustrates the worst case and mean value of IRR, as well as the notch position drift and the IRR distribution between 50dB and 90dB.

PPF stages number	*Mean value of the IRR*	*Worst case IRR*	*Notch drift*	*Standard deviation σ*	*IRR distribution between 50dB and 90dB*
3	62dB	51dB	405MHz	9.11	85%
4	72dB	57dB	306MHz	10.05	95%
5	87dB	64dB	317MHz	11.95	97%

Table 3. Monte Carlo simulation results of multi-stage RF PPFs: mean value and worst case IRR, notch position drift and IRR distribution between 50dB and 90dB

The obtained results confirm that increasing the stages number increases the mean value of the IRR on the desired bandwidth. It can be noted that the higher is the PPF stages number, the lower is the PPF immunity to mismatch effects, given that the distribution becomes wider and the standard deviation $σ$ increases from 9.11 to 10.05 and 11.95 for three-stage, four-stage and five-stage RF PPFs respectively. This is due to the components and connections growth in the design, inducing, at the same time, an expansion of its area.

Let us consider a unit surface S_0 of a one-stage RF PPF. Since in the RF domain the size of our PPF components is almost identical, we can suppose that an n-stage PPF has a surface of $n.S_0$. Thus, a compromise can be made while designing PPFs depending on the system specifications. For example, a 60dB image rejection will cost $3.S_0$ with a standard deviation of 9, while a roughly 85dB image rejection will cost $5.S_0$ with a standard deviation of 12.

6. Parasitics analysis and line modeling

Since the implementation of RC polyphase filter on integrated circuit engenders parasitic capacitance to the substrate and at the output nodes, special attention must to be paid on the parasitic capacitance and loading capacitance effects. In Fig.12 we model a simplified equivalent circuit of a two-stage RC PPF with parasitic capacitance to substrate (Cp1, Cp2, Cp3) and load capacitance (a part of Cp3).
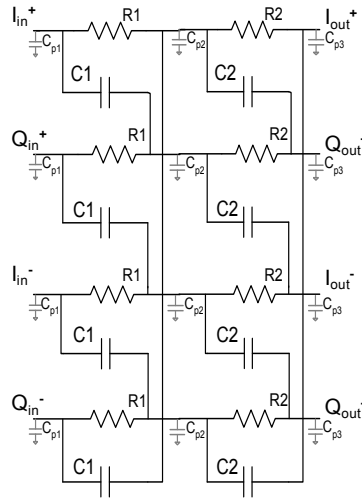
Fig. 12. Equivalent circuit of the two-stage RC PPF with parasitic capacitance

In this case, the transfer functions of one-stage and two-stage RC polyphase filter are given respectively as follows

$$H_{p1}(j\omega) = \frac{1 - \omega R_1 C_1}{1 + \omega R_1 (C_1 + C_{p2})}$$

(18)

$$H_{p2}(j\omega) = \frac{(1 - \omega R_1 C_1)(1 - \omega R_2 C_2)}{1 - \omega^2 R_1 R_2 \left[C_1 C_2 + C_1 C_{p2} + C_2 (C_{p2} + C_{p3}) + C_{p2} C_{p3} \right] + j\omega \left[R_1 (C_1 + 2C_2 + C_{p2} + C_{p3}) + R_2 (C_2 + C_{p3}) \right]}$$

(19)

It can be noted from (18) and (19) that the parasitic capacitances do not change the zero positions $1/2\pi R_1 C_1$ and $1/2\pi R_2 C_2$ of $H_{p1}(j\omega)$ and $H_{p2}(j\omega)$. Simulation results of frequency response including parasitic capacitances depict that the gain drops for high frequency domain when the parasitic capacitance values increase (Yamaguchi et al, 2003).

Furthermore, properly arranging the components and optimally sizing the connections are necessary to guarantee an equilibrated parasitic repartition in the circuit, which can conserve the symmetrical structure of passive polyphase filter. The major loss and parasitic capacitance contributions in connections are considered in order to obtain better filter performance. In fact, loss in a conductor can be generally described by the following equation

$$R = \frac{\rho_{film} \cdot L}{t \cdot W}$$

(20)

where ρ_{film} is the thin film resistivity of the metal, t is the metal thickness, and L and W are the trace length and width, respectively. Therefore, loss can be minimized by using metals with very low resistivity, increasing the cross sectional area of the trace ($t.W$), or reducing the overall trace length. Besides, the metal of connection is isolated from the semiconductor substrate (typically at ground potential) by one or more dielectric layers used to separate interconnect layers (inter-metal dielectrics). This creates a parasitic shunt capacitor that can be approximated by the following equation

$$C = \frac{A.\varepsilon}{d} \qquad (21)$$

where A is the total area of the metal traces, ε is the permittivity, and d is the thickness of the dielectric. The parasitic capacitance decreases with high metals levels, but at the same time this will increase the parasitic resistance because of stacking the different "via" resistivities. Hence, designers must balance both the parasitic shunt capacitance and conductor loss when selecting a conductor dimensions and metals levels.

Characterization and modeling of the interconnection lines have been performed to improve their properties. The equivalent network line model between two ports used in this study is shown in Fig.13(a). First, the line parameters have been extracted with electromagnetic simulations (*HFSS*[TM]). Then, the correspondent line models have been specified and inserted inside the polyphase filter design at the main sensitive points and simulated with the *Agilent ADS®* tool. Calibration of the additional parasitics allows their allocation symmetrically along the design, since their total elimination is not possible. This study has demonstrated that lines with different shapes give the same filter response (IRR and bandwidth) provided that the interconnect lengths in respectively I/Q paths are equalized. It is caused by the fact that this will balance the parasitic interconnect resistance in each branch. For example, serpentine and bus shapes could be used simultaneously for the parallel interconnections. By adjusting the height of serpentine, the wire length in the branches of the PPF may be equalized while keeping the same number of corners (Fig 13(b)).

(a)   (b)
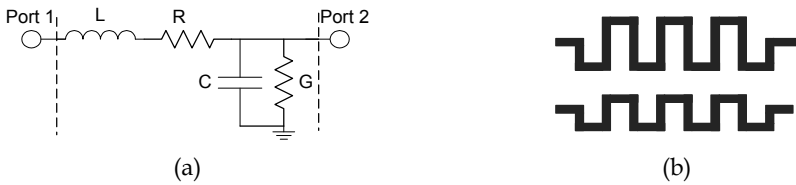
Fig. 13. (a) Equivalent network line model between different levels of interconnect. (b) Considerations of interconnect to balance parasitics in polyphase filter branches

Besides, the inaccuracy of resistors and capacitors, due to Si substrate parasitic effect, causes quadrature phase imbalance. To overcome this problem it is possible to make the polyphase filter tunable so as to compensate the phase imbalance. The tunable phase can be used to improve image rejection or moderate I/Q phase error in direct conversion or low-IF receivers. For instance, varactor-based tunable polyphase filters on Si have been implemented at 5GHz (Sanderson et al, 2004). Another technique to solve RC inaccuracy of PPF is to use InGa/GaAs heterojunction bipolar transistor which has a very good frequency response but which remains expensive (Meng et al, 2005). In addition, in the RF front-end receiver, the input large parasitic capacitances of the following double quadrature mixer degrade the loss of the RF polyphase filter. To overcome this problem, on-chip spiral inductors are inserted at the output of the RF PPF in (Kim & Lee, 2006) and then tune out the total input parasitic capacitances of the double quadrature mixer.

In our design, a new polyphase filter implementation (shown in Fig.14) is proposed to balance the bandwidth variation due to mismatches in a symmetrical structure. It consists on the RC basic passive polyphase network, adding up active resistors implemented with MOS transistors. It is known that the R_{on} of the MOS transistor is function of its dimensions

and of the grid voltage (VG). Thus, with an external tuning of VG, the value of R_{on}, and then the PPF resistor value and the notches, can be adjusted independently. Consequently, that gives a tuning characteristic to the filter bandwidth, and can be applied to synthesize multi-standards application filters. The MOS transistor dimensions are chosen to have the adequate calibration of the bandwidth dispersion. Using these MOS active resistors possibly adds nonlinearity to the PPF design, and then other active resistor realizations, such as parallel-MOS and double-MOS differential resistor, with better linearity performance, have been proposed (Allen & Holberg, 2002).
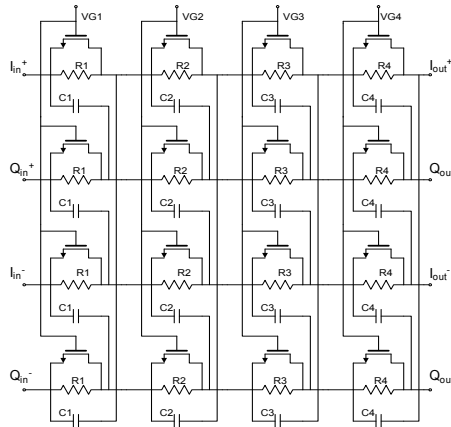
Fig. 14. Four-stage voltage tunable RC polyphase filter structure

7. Layout techniques

In addition, while components with large areas decrease the impact of mismatch, the parasitic capacitance and resistance can have a much larger effect on output imbalance. Minimization of these parasitics requires careful attention to layout symmetry. The parasitic extraction procedure, performed with the *Star-RCXT* tool of *Synopsys*, shows that most extracted parasitics are set in the interconnection network. Interconnects present electrical losses that need to be taken into account during layout and then during performances estimation. It is clear that, on the circuit, the inner traces see parasitic capacitance from the left and right, while the outer traces only see parasitic capacitance from one side. Hence, weaving the traces gives each path the same total distance spent as both an inner and an outer trace. To equalize the parasitic effect of overlapping traces, a grid of vertical and horizontal running interconnects has been laid out. Moreover, two parallel signal lines are placed far enough apart so that the interline capacitance is negligible.

Furthermore, a judicious choice of metal level and interconnection drawing is necessary. In fact, using high level of metallization engenders low parasitic capacitance but gives high parasitic resistance. Thus, depending on the device sensibility and on the required matched components, the metal level is chosen. For example, in low-loss applications, the metal 6 is the most suitable (in 0.13µm CMOS technology) since it is the thickest one and has less capacitance. The number of "vias" used for interconnects is also significant in leading to equilibrated parasitics, especially in the case of RF passive polyphase filters. These vias give

high contact resistance that can be almost equal to the filter resistance. Table 4 presents some extraction results of a line connection between a resistor and a capacitor having different metal levels with different vias number. It shows that increasing the number of vias does not change the parasitic capacitance, but decreases the parasitic resistance. It is due to putting the contact resistance of each via in parallel and then lowering the equivalent resistance. Therefore, connections in the radio frequency PPF have to use great number of vias to minimize their parasitic effect.

Metal level of the line	Via number	Parasitic capacitance (fF)	Parasitic Resistance (Ω)
Metal 2	1	14.5	19.11
	2	14.54	11
	3	14.56	7
Metal 3	1	14.13	19.11
	2	14.15	13.49
	3	14.18	10.82

Table 4. Extraction results of a line connection (W=2.5-µm/L=5-µm) with different metal levels between a Polysilicon-resistor and a MIM-capacitor in 0.13µm CMOS technology

Total equilibrated interconnects drawing is hard to obtain in the case of PPF. However, owing to the symmetry of the PPF stages, the parasitic modelling and extraction procedures illustrate that ensuring the same drawings between I and Q paths is sufficient to guarantee same matching and same performances as in the case of an ideal structure (with same drawings for the four PPF paths), and then, that may loosen the constraints of design techniques.

In addition to designing a symmetrical circuit, further layout techniques have been used to assure highly matched devices, as shown below

- To reduce the sensitivity of the device to process biases, resistors are made same width and capacitors consider same area-to-periphery ratios.
- Dummy resistors are added to either border of an array of matched resistors to guarantee uniform etching. Dummies should be electrically connected to ground (or to other low-impedance node) to avoid electrostatic modulation and floating diffusions. Moreover, the metal overlapping the active area of resistors can lead to metallization-induced mismatches. Thus, the "folded-out" interconnection (Fig.15(a)) produces better matching than the "folded-in" interaction (Fig.15(b)).
- Stress has an impact upon silicon since it is piezoresistive. One of the most known techniques for reducing stress-induced mismatches is the *common-centroid layout*. It arranges segments of matched devices along one dimension. For example, if we consider two devices (A and B), each composed of two segments, the possible patterns are shown in Fig.15(c). The pattern ABBA has an axis of symmetry that divides it into two mirror-image halves (AB and BA). It requires dummies since segments of A occupy both ends of the array. The pattern ABAB, with interdigitated resistors, haven't common axis of symmetry and needs dummies as well as the ABBA pattern. Thus, the pattern ABAB lets stress-induced mismatches on devices and consequently it should be avoided (Hastings, 2006).

- Thermoelectric effects cannot be eliminated with the common-centroid layout in the case of an array of resistors, because they arise from differences in temperature between the ends of each resistor segment. The thermoelectric potentials of individual segments can be cancelled by reconnecting them as shown in Fig.15(d). The resistor should have an even number of segments, half connected in one direction and half connected in the other.
- Electrostatic interactions cause variations in resistors and capacitances. Thus, matched resistors with same values can belong to a common tank (or N-wells). If resistors have different values, they should be divided into segments of equal values, and each segment must reside in its own independently biased tank. In addition, wires that do not connect matched resistors should not cross them, because they may capacitively couple noise into the resistor and the electric field between the wire and the resistor can modulate the conductivity of the resistance material. The *electrostatic shielding* (or *Faraday shielding*) is a technique that can isolate a resistor from the influence of overlying leads and gives shielding against capacitive coupling (Hastings, 2006).
- To avoid electromigration between signals, I and Q paths are separated with a grounded bus.
- Size, orientation and temperature stress of MOS transistors influence their matching. A better matching is obtained when transistors are oriented along the same crystal axis in the same direction because of the stress-induced mobility variations. They should also be placed in close proximity even next to one another in order to facilitate common-centroid layout.

| (a) | (b) | (c) | (d) |
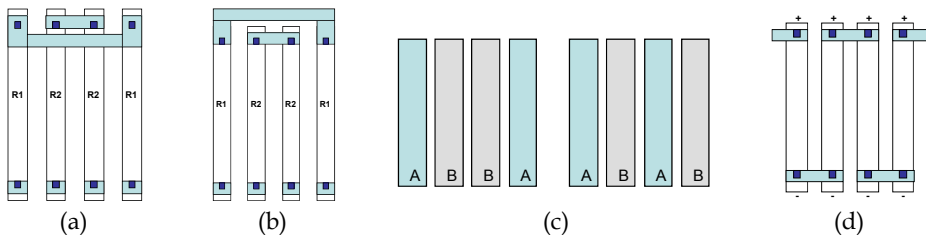
Fig. 15. Resistor array interconnection in (a) "folded-in" and (b) "folded-out" styles. (c) Examples of common-centroid arrays. (d) Proper connection of resistor segments cancelling the thermoelectric

8. PPF Design methodology

As analyzed previously, component mismatch, process tolerances and parasitic effects must be considered in the design of CMOS PPFs to accomplish a robust design. We propose a design methodology dedicated to PPFs as shown in Fig.16. Such top-down design methodology is a structured approach to design PPFs operating from wide frequency range and which can satisfy high performances in terms of IRR (about 60dB) from wide frequency range (1MHz to 5GHz).

This PPF design methodology can be arranged into considerations first in the system requirements, then in the schematic design and next in the layout view. Thus, starting out from target specifications and constraints in terms of IRR, application bandwidth, cost and consumption, we can summarize the design flow as the guidelines below

- Accomplishing analytical calculations and modeling to quantify the component mismatch and parasitic elements effects and to focus on the resulting PPF response to phase and gain imbalances.
- Fixing the number of stages needed for the polyphase filter according to the bandwidth to be covered and the desirable image rejection amount.
- Equally placing the notches on the frequency domain with growing impedance while traversing the filter stages to lower losses and noise figure.

If the cascade filter loss is still too large, we move on changing the component type as well as calibrating its parameters, even as inserting inter-stage buffers to preserve signal dynamic range within the polyphase filter. After adjusting the losses into the PPF, we fulfill statistical simulations to longer analyze the component mismatch.

- Optimal sizing of the PPF components in terms of electrical value and dimensions. The matching quantities needed between resistors and capacitors determine the physical area of the filter.

If in the schematic simulation, the target specification cannot be met, we move on to the component resizing procedure and deduce the compliance with the required constraints. After completing the schematic design, we carry out the physical layout design.

- Modeling the interconnection lines and performing electromagnetic simulation to deduce their parameters; and then inserting them in the PPF design to maximize its immunity to the non idealities.
- Designing the layout taking into account the parasitic elements: the conductor loss of the interconnect metal creates parasitic resistance, and the dielectric between the traces and the substrate or between two overlapping traces creates parasitic capacitance. Layout which creates equal parasitics for each path through the polyphase is necessary to minimize the imbalance and maintain the symmetry.
- Using dummies around the matched components to reduce the boundary effects and on-chip shielding to isolate the PPF design from the unwanted substrate noise coupling. The electromigration is minimized with a ground separation between the I and Q signals. A judicious choice of the metal level and number of contacts or vias is also necessary.
- Post layout simulating the PPF with the extracted coefficients. In this extraction method, parasitics between neighboring components, wires and parasitics to the substrate are extracted. In this way, we can provide realistic simulation results before manufacturing the circuit.

If the target specifications required by the application are not yet satisfied, we go back to the parasitics minimization procedure and post-layout simulation (PLS) until assuring them. Then, we finish the design.
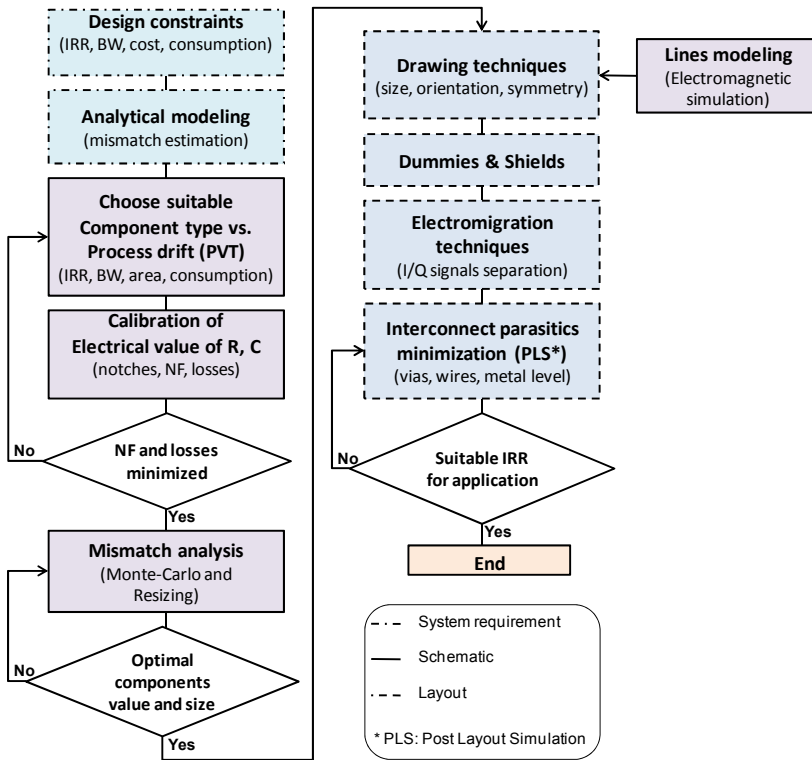
Fig. 16. High performance PPF design planning flow

9. PPF implementation

The proposed design methodology has been validated with some test-cases in full CMOS process. For instance, Fig.17 shows the layout of a four-stage RF tunable PPF (rf. Fig.14) designed to work around 5GHz, and fabricated in 0.13-µm CMOS technology. It occupies a die area of 310 x 83 µm² without test pads.
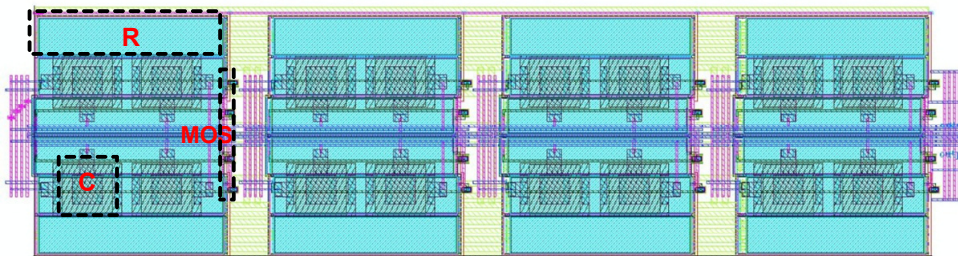

Fig. 17. Layout of the 5GHz four-stage tunable PPF: 310 x 83 µm² without test pads

The frequency response of the implemented 5GHz tunable PPF is depicted in Fig.18. It shows that the variation of the control grid voltage of MOS resistors enables the tuning of the PPF bandwidth by 1GHz while conserving an IRR almost steady around 75dB. Then, this proposed tuning characteristic can be applied to multi-standard applications, or used to compensate for the bandwidth drift due to mismatches.
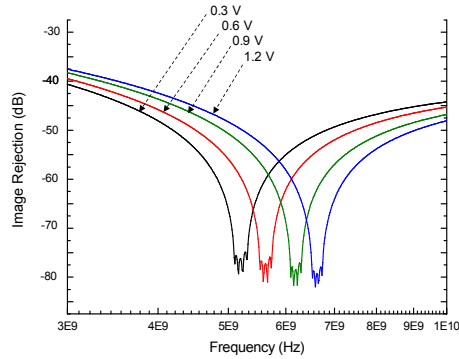
Fig. 18. Frequency responses of the 5GHz tunable polyphase filters using different control grid voltages

A chip photo of the fabricated chip is shown in Fig.19. It occupies 815 x 319 µm² with test pads. On-chip polysilicon resistors have been added to recombine the four outputs of the PPF in order to avoid the inaccuracy of the external hybrid couplers and to facilitate the measurement procedure. Thus, a differential output is obtained and can be measured easily with active probes.
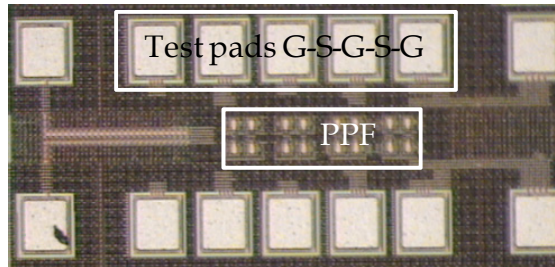
Fig. 19. Die micrograph of the fabricated PPF test chip in 0.13-µm CMOS technology
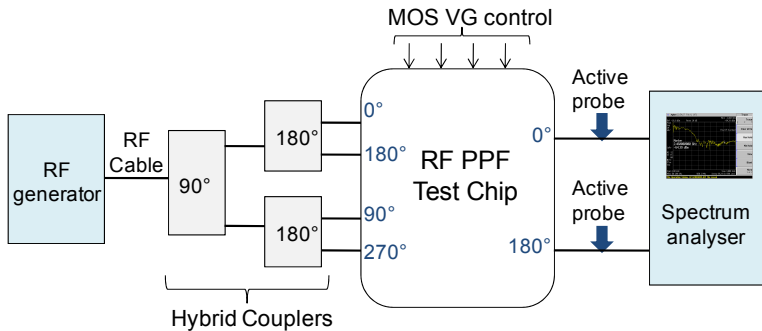
Fig. 20. Diagram of the PPF measurement setup

A diagram of the measurement setup for test of the CMOS PPF is illustrated in Fig.20. On-wafer RF measurements can be performed since balanced G-S-G-S-G pads (*G* for ground and *S* for signal) are used. The four input phases of the filter are generated by using a signal generator and wideband passive hybrid couplers. The measurement procedure is being processed to validate the obtained simulation results.

10. Conclusion

Wireless communication bands and services are proliferating, resulting in a great development of standards and in an enhanced need for integrated circuits. In this paper, it is demonstrated that techniques for image rejection have been constantly evolving in recent years because of this tremendous success of wireless products. Among the various techniques, the polyphase filters might become the choice for future image rejection scheme, thanks to its promising performances and to the semiconductor process advances. An analytical approach of RC polyphase filters as well as a study of components mismatch and non-ideality impact on the IRR degradation have been presented in this paper. That leads us to propose a design methodology dedicated to passive polyphase filters (PPFs), taking into account optimum component sizing, lines modeling and layout symmetry and matching. This method has been validated with some test-cases in full CMOS technology and allows attaining high image rejection (about 60dB) from wide frequency range (1MHz to 5GHz).

In addition, the wireless services have different carrier frequencies, channel bandwidths, modulation schemes, data rates, etc., which motivates the industry to look for multi-standard and multi-band devices. In this paper, a tunable polyphase filter structure has been proposed, which can be applied to synthesize multi-standard application filters. This tuning characteristic can be also used to compensate for the bandwidth drift due to mismatches.

11. References

Allen, P. E. & Holberg, D. R. (2002). *CMOS Analog circuit design* (éd. 2nd). New York: Oxford University Press.

Andreani, P.; Mattisson, S. & Essink, B. (2000). A CMOS gm-C polyphase filter with high image band rejection. *Proceedings of the European Solid State Circuits Conference*, 244-247.

Andreani, P. & Mattison, S. (2002). On the use of Nauta's transconductor in low-frequency CMOS gm-C bandpass filters. *IEEE Journal of Solid State Circuits*, 37, 114-124.

Banu, M.; Wang, H.; Seidel, M.; Tarsia, M.; Fischer, W.; Glas, J.; Dec, A. & Boccuzzi, V. (1997). A BiCMOS Double low-IF receiver for GSM. *Proceedings of IEEE Custom Integrated Circuits Conference*, 521-524.

Behbahani, F.; Kishigami, Y.; Leete, J. & Abidi, A. A. (1999). CMOS 10 MHz IF downconverter with on-chip broadband circuit for large image suppression. *Symposium on VLSI circuits*, 83-86.

Behbahani, F.; Weeguan, T.; Karimi-Sanjaani, A.; Roithmeier, A. & Abidi, A. A. (2000). A broadband tunable CMOS channel select filter for a low-IF wireless receiver. *IEEE Journal of Solid State Circuits*, 35 (4), 476-489.

Behbahani, F.; Kishigami, Y.; Leete, J. & Abidi, A. A. (2001). CMOS mixers and Polyphase filters for large image rejection. *IEEE Journal of Solid State Circuits*, 36 (6), 873-887.

Behbahani, F.; Firouzkouhi, H.; Chokkalingam, R.; Delshadpour, S.; Kheirkhahi, A.; Nariman, M.; Conta, M. & Bhatia, S. (2002). A fully integrated low-IF CMOS GPS radio with on-chip analog image rejection. *IEEE Journal of Solid State Circuits*, 37 (12), 1721-1727.

Behzad, A.; Lin, L.; Shi, Z.M.; Anand, S.; Carter, K.; Kappes, M.; Lin, E.; Nguyen, T.; Yuan, D.; Wu, S.; Wong, Y.C.; Fong, V. & Rofougaran, A. (2003). Direct conversion CMOS transceiver with automatic frequency control for 802.11a wireless LANs. *Solid State Circuits Conference ISSCC*, 1, 356-499.

Carta, C.; Vogt, R. & Bachtold, W. (2005). Multiband monolithic BiCMOS low-power low-IF WLAN receivers. *IEEE Microwave and Wireless Components Letters*, 15 (9), 543-545.

Chian, L. et al. (2007). Design of tunable polyphase filter using MOSFET. *Proceedings of the International Symposium on Integrated Circuits*, 349-352.

Chou, S. & Lee, H. (2007). Effect of phase mismatch on image rejection in Weaver architecture. *IEEE Microwave and Wireless components Letters*, 17 (1), 70-72.

Chou, C. & Wu, C. (2005). The design of wideband and low-power CMOS active polyphase filter and its application in RF double-quadrature receivers. *IEEE Trans. on Circuits and Systems I: Regular Papers*, 52 (5), 825-833.

Crols, J. & Steyaert, M. (1995). A single chip 900 MHz CMOS receiver front-end with a high performance low-IF topology. *IEEE Journal of Solid State Circuits*, 30 (12), 1483-1492.

Crols, J. & Steyaert, M. (1998). Low-IF topologies for high-performance analog font ends of fully integrated receivers. *IEEE Trans. on Circuits and Systems II*, 45, 269-282.

Der, L. & Razavi, B. (2003). A 2 GHz CMOS Image-reject receiver with LMS calibration. *IEEE Journal of Solid State Circuits*, 38 (2), 167-175.

Elmala, M. & Embabi, S. (2004). Calibration of phase and gain mismatches in Weaver image-reject receiver. *IEEE Journal of Solid State Circuits*, 39 (2), 283-289.

Fang, S. et al. (2005). An image rejection down converter for low-IF receivers. *IEEE Trans. on Microwave Theory and Techniques*, 53 (2), 478-487.

Galal, S. H.; Ragaie, H. F. & Tawfik, M. S. (2000). RC sequence asymmetric polyphase networks for RF integrated transceivers. *IEEE Trans. on circuits and Systems II*, *47* (1), 1127-1134.

Galal, S. H. & Tawfik, M. S. (1999). On the design and sensitivity of RC sequence asymmetric polyphase networks in RF integrated transceiver. *Proceedings of IEEE International Symposium on Circuit and System*, *2*, 593-597.

Gingell, M. (1971). *Brevet n° 3,559,042; 3,618,133.* US patents.

Harrison, J. (2002). 350 MHz opamp-RC filter in 0.18 µm CMOS. *Electron Letters*, *38* (6), 259-260.

Hastings, A. (2006). *The art of analog design.* New Jersey: Prentice Hall, 2nd Edition.

Huang, Q.; Orsatti, P. & Piazza, F. (1999). GSM transceiver front-end circuits in 0.25 µm CMOS. *IEEE Journal of Solid State Circuits*, *34* (3), 292-303.

Khvedelidze, B. (2001). Hilbert transform. Dans M. Hazewinkel, *Encyclopaedia of Mathematics.* Netherlands: Kluwer Academic Publishers.

Kim, C. & Lee, S. (2006). A 5.25 GHz image rejection RF front-end receiver with polyphase filters. *IEEE Microwave and Wireless Components Letters*, *16* (5), 302-304.

Komoriyama, K; Yashiki, M.; Yoshida, E. & Tanimoto, H. (2007). A very wideband fully balanced active RC polyphase filter based on CMOS inverters in 0.18 µm CMOS technology. *Proceedings of Symposium on VLSI Circuits*, 89-99.

Lee, S.; Jung, K.; Kim, W.; Ryu, H. & Song, W. (1998). A 1 GHz image-rejection downconverter in 0.8 µm CMOS technology. *IEEE Trans. on Customer Electronics*, *44* (2), 235-239.

Lee, T.H.; Samavati, H. & Rategh, H.R. (2002). 5 GHz CMOS Wireless LANs. *IEEE Trans. on Microwave and Theory Techniques*, *50* (1), 268-280.

Lerstaveesin, S. & Song, B. (2006). A complex image rejection circuit with Sign detection only. *IEEE Journal of Solid State Circuits*, *41* (12), 2693-2702.

Long, J. R. (1996). A narrowband radio receiver front-end portable communications applications. *Ph.D dissertation* . Carleton University.

Long, J. R. & Maliepaard, M. (1999). A 1V 900 MHz image-reject downconverter in 0.5 µm CMOS. *IEEE Custom Integrated Circuits Conference*, 665-668.

Macedo, J. A., & Copeland, M. A. (1998). A 1.9 GHz Silicon Receiver with monolithic image filtering. *IEEE Journal of Solid State Circuits*, *33* (3), 378-386.

Maligeorgos, J. & Long, J. (2000). A 2V 5.1-5.8 GHz Image-reject receiver with wide dynamic range. *IEEE Solid-State Circuits Conference ISSCC*, 322-323.

McNutt, M.J.; LeMarquis, S. & Dunkley, J.L. (1994). Systematic capacitance matching errors and corrective layout procedures. *IEEE Journal of Solid State Circuits*, *29* (5), 611-616.

Meng, C.C.; Sung, D.W. & Huang, G.W. (2005). 5.2 GHz GaInP/GaAs HBT double-quadrature downconverter with polyphase filters for 40 dB image rejection. *IEEE Microwave and Wireless Components Letters*, *15* (2), 59-61.

Mikkelsen, J. H. (1998). Evaluation of CMOS front-end receiver architectures for GSM handset applications. *IEEE Symp. Communications Systems and Digital Signal Processing*, 164-167.

Montemayor, R. & Razavi, B. (2000). A self-calibrating 900 MHz image reject receiver. *Proceedings of ESSCIRC*, 292-295.

Razavi, B. (1996). Challenges in portable RF transceiver design. *IEEE Circuits and Devices Magazine*, *12*, 12-25.

Razavi (a), B. (1997). *RF Microelectronics.* New Jersey: Prentice Hall.

Razavi (b), B. (1997). Design considerations for direct conversion receivers. *IEEE Trans. Circuits and Systems II*, *44* (6), 428-435.

Razavi, B. (2001). A 5.2 GHz CMOS receiver with 62 dB image rejection. *IEEE Journal of Solid State Circuits*, *36* (5), pp.810-815.

Rudell, J. C.; Ou, J.-J.; Cho, T.B.; Chien, G.; Brianti, F.; Weldon, J.A. & Gray, P.R. (1997). A 1.9 GHz wideband IF double conversion CMOS receiver for cordless telephone applications. *IEEE Journal of Solid State Circuits*, *32* (12), 2071-2088.

Samavati, H.; Rategh, H.R. & Lee, T.H. (2000). A 5 GHz CMOS Wireless LAN receiver front end. *IEEE Journal of Solid State Circuits*, *35* (5), 765-772.

Samavati, H., & al., e. (2001). A fully integrated 5 GHz CMOS Wireless LAN receiver. *Proceedings of ISSCC*, 208-209.

Sanderson, D. et al. (2004). A 5-6 GHz polyphase filter with tunable I/Q phase balance. *IEEE Tansactions on Microwave and Wireless Components Letters*, *14* (7), 364-366.

Spence, R. & Soin, R. (1997). *Tolerance design of electronic circuit.* New York: Addison-Wesley.

Steyaert, M. & Crols, J. (1994). Analog integrated polyphase filters. *Proceedings of the workshop on advances in Analog Circuit Design*, p.18.

Steyaert, M.; Janssens, J.; de Muer, B.; Borremans, M. & Itoh, N. (2000). A 2V CMOS cellular transceiver front end. *IEEE Journal of Solid State Circuits*, *35* (12), 1895-1907.

Sun, N., Lee, H.S. & Ham, D. (2008). Digital background calibration in pipelined ADCs using commutated feedback capacitor switching. *IEEE Trans. circuits and Systems II*, *55* (9), 877-881.

Tetsuo, Y. (1995). Polyphase network calculation using a vector analysis method. *QEX including Communications Quarterly*, 9-15.

Thandri, B. & Silva-Martinez, J. (2003). A robust feedforward compensation scheme for multistage operational transconductance amplifiers with no Miller capacitors. *IEEE Journal of Solid State Circuits*, *38* (2), 237-243.

Tillman, F. & Sjoland, H. (2005). A polyphase filter based on CMOS inverters. *Proceedings of the 23rd IEEE NORCHIP Conference*, 12-15.

Tsividis, Y. (1994). Integrated continuous-time filter design - An overview. *IEEE Journal of Solid State Circuits*, *29* (3), 166-153.

Ün (a), M. (2004). True polyphase filter section implemented with CCIIs. *Trans.on Engineering, Computing and Technology*, 160-163.

Ün (b), M. (2004). Performance analysis of RC-active polyphase filter section implemented with OPAMPs. *Trans. on Engineering, Computing and Technology*, 353-356.

Valkama, M. & Renfors, M. (2000). Advanced DSP for I/Q imbalance compensation in a low-IF receiver. *IEEE International Conference on Communications*, 2, 768-772.

Van Zeijl, P., Eikenbroek, J.W.; Vervoort, P.P.; Setty, S.; Tangenberg, J.; Shipton, G.; Kooistra, E.; Keekstra, I. & Belot, D. (2002). A Bluetooth radio in 0.18 μm CMOS. *Journal of Solid State Circuits*, *37* (12), 1679-1687.

Wu, C. & Chou, C. (2004). A 5 GHz CMOS double quadrature receiver front-end with single-stage quadrature generator. *IEEE Journal of Solid State Circuits*, *39* (4), 519-521.

Wu, S. & Razavi, B. (1998). A 900 MHz/1.8 GHz CMOS receiver for dual-band applications. *IEEE Journal of Solid State Circuits*, *33* (12), 2178-2185.

Wu, C. & Chou, C. (2003). A 5 GHz CMOS double-quadrature receiver for IEEE 802.11a applications. *Symposium on VLSI Circuits Digest of Technical Papers*, 149-152.

Xu, J.; Chen, J. & Zheng, J. (2001). Design of Weaver Topology. *Electronic Letters*, *37* (18), 1133-1135.

Yamaguchi, N., Kobayashi, H., Kang, J., Niki, Y., & Kitahara, T. (2003). Analysis of RC polyphase filters - High-order filter transfer functions, Nyquist charts, and parasitic capacitance effects. *IEIC Technical Report*, *102* (572), 29-34.

Fully Integrated CMOS Low-Gain-Wide-Range 2.4 GHz Phase Locked Loop for LR-WPAN Applications

Wenceslas Rahajandraibe, Lakhdar Zaïd and Fayrouz Haddad
IM2NP – University of Provence
Marseille - France

1. Introduction

The last decade has been marked by a rapid growing of wireless market and this phenomenon trends to accelerate in future years. This market serves different demands in wireless applications for cellular phones, wireless local area networks (WLAN), wireless personal area networks (WPAN), phased array RF systems, and other emerging wireless communication such as wireless body area network (WBAN), radar, and imaging applications operating in a very wide frequency range: few MHz up to 100GHz (ITRS, 2007). The introduction of digital signal processing inside communication systems constitutes one of the main reasons of this growth. This digital revolution results from research and development related to high performance CMOS technologies, coming with lower cost than classical bipolar technology and allows the integration of complex digital and analog function on the same chip. Today, digital evolution and the market flight of mobile communications lead to several changes in the analog part of the radio-frequency (RF) front end of transceivers (interface between antenna and digital modem). The need for RF front end to detect very weak signal (few µV) at very high frequency (~GHz) and in the same time to be able to transmit high power signal (few Watts) requires high performance analog circuits such as filters, amplifiers, mixers and oscillators. Historically, RF communications was reserved to military uses where the performance predominated without real cost constraints. The introduction of wireless communication in commercial and public domain where cost reduction is the leitmotiv has leaded the analog part to be the most critical part of current and future RF systems (Chen, 2000).

2. Evolution of LR-WPAN: Standardization

Coming with rapid developments of information technology in the 1980s, laptops have begun to be used elsewhere than as part of the office. With the accession of the Internet in 90's, mobility has become problematic: strong demand appeared to allow connecting to the internet everywhere. The emerged solution was to connect computers to each other by the

way of radio wave rather than wire, resulting to wireless local area network (WLAN). WLAN requires a fixed access point that can connect multiple mobile stations.

The dramatic rise of the demand and application fields has conducted to standardization. It defines an interface between "client" and "access point" in the wireless network by specifying both the physical layer (PHY) and the software layer (or MAC: Medium Access Control). The goal is to ensure the interoperability of data networking, the security services and a range of wireless home and building control solutions. This will assure consumers to buy products from different manufacturers with confidence that the products will work together (ZigBee Alliance). Working group is formed to create different standards according to their characteristics: distance of coverage, data-rate, communication protocol, etc. The IEEE 802.15 working group relates wireless personal area network (WPAN) which focuses low-cost, low power, short range and very small size circuit. There are three classes of WPAN according to data rate, battery life, and quality of service (QoS). The high data rate PAN (IEEE 802.15.3) is suitable for multi-media applications (streaming) that require very high QoS. Medium rate WPANs (IEEE 802.15.1/Blueetooth) will handle a variety of tasks ranging from cell phones to PDA communications and have QoS suitable for voice communications. The low rate WPANs (IEEE 802.15.4/LR-WPAN) is intended to serve a set of industrial, residential and medical applications with very low power consumption and cost requirement not considered by the above WPANs and with relaxed needs for data rate and QoS.

ZigBee standard is one of existing LR-WPAN. It is expected to provide low cost and low power connectivity for equipment that needs battery life as long as several months to several years but does not require data transfer rates as high as those enabled by Bluetooth. In addition, ZigBee can be implemented in mesh networks larger than that of Bluetooth. ZigBee compliant wireless devices are expected to transmit 10-75 meters, depending on the RF environment and the power output consumption required for a given application, and will operate in the unlicensed RF worldwide (2.4GHz global, 915MHz Americas or 868MHz Europe). The data rate is 250kbps at 2.4GHz, 40kbps at 915MHz and 20kbps at 868MHz. (ZigBee Alliance)

3. Wireless communications: transceiver circuit challenges

Each wireless transceiver, responds to its proper characteristics and constraints according to the application, in order to achieve an efficient transmission of the data without altering neighbor transceivers. Among the main characteristics, one can note the maximum distance of coverage, the number of the communication channels, the value of the carrier frequency, the power level of the transmitted signal, the bit-error-rate (BER), the noise and so on. Mobile applications are subjected to many constraints, namely the circuit cost, the autonomy of the battery, the interoperability with other applications, etc.

The operating characteristics of the transceiver can be derived from the standard definition; however, hard constraints related to the system architecture, the power and the cost constitute real challenges for current and future wireless communications. These performances depend both on the quality and the cost of the technology used to implement the design and on the design solution adopted to meet the standard as well as the given specification requirement.

3.1 Technology consideration

The feasibility of many wireless products mainly depends on the intrinsic performances of the technology used in radio-frequency (RF) and analog/mixed-signal (AMS) which can be divided to four categories depending on the field of applications. Compound III-V semiconductors (GaAS, InP, etc.) have traditionally dominated the millimeter wave spectrum over the past several decades. However, today, with the drive to low-cost high-volume applications such as auto radar, along with scaling to sub-100nm dimensions, devices implemented with Si and SiGe are rapidly moving up to frequencies that were once the exclusive domain of the III-Vs. CMOS, BiCMOS and SiGe for heterojunction bipolar transistor are the most adopted process, while implementing monolithic system-on-chip (SoC) and intellectual property (IP) for wireless applications.

Generally, the choice criterion of the technology is driven by cost, frequency bands, power consumption, functionality, volumes of product and standards and protocols. Today, BiCMOS in cellular transceivers has the biggest share in terms of volume compared to CMOS. But, the opposite may occur in the future as evident by the expanding wireless local area network (WLAN) connectivity market that is dominated by CMOS transceivers (ITRS, 2007). CMOS process is mainly used to implement on chip digital circuits since it allows high integration density with a lower cost than any other processes. The size reduction and the process refinement of CMOS devices allowed increasing the transition frequency and the operating frequency of RF and analog/mixed signal circuits. Several wireless transceiver designs (GSM, DECT, DCS1800, etc.) have taken benefits of this feature and have been efficiently realized with CMOS technology (Mikkelsen, 1998). However, scaling down the gate size comes with supply voltage reduction, penalizing the voltage dynamic, signal-to-noise ratio, and linearity. Additional process step is then required during the fabrication for higher voltage supply increasing the cost. Figure 1 depicts some examples of wireless applications fully implemented in CMOS technology as a function of operating frequency (Crols & Steyaert, 1995).
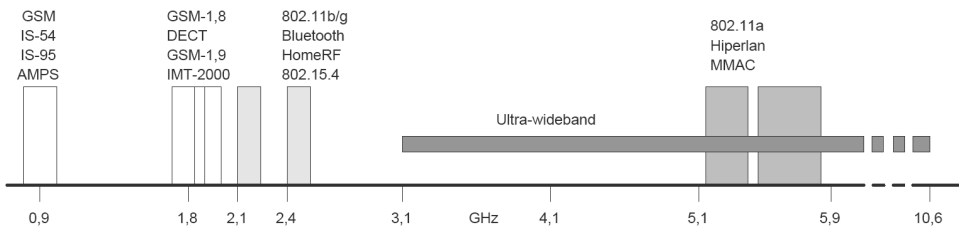
Fig. 1. Repartition of some wireless applications versus operating frequencies

3.2 Wireless communication design challenges

The constraints imposed by Bluetooth or wireless fidelity (WiFi) standards in terms of data-rate, channels spacing and access method (CDMA) are not compatible with the design objectives related to the achievement of low-cost products such as LR-WPAN. IEEE 802.15.4 aims to define a production cost of chips < \$2, with substantial autonomy battery life (>1 year). In practice, this hard consumption requirement imposes the system to standby for 99.9% of battery lifetime. Note that this network will coexist with other networks operating in the same frequency band (Bluetooth, WiFi, etc.). Thus, the fact that it operates only 1% of the time makes LR-WPAN system little disruptive of other networks.

Despite recent advances in terms of power consumption: dedicated circuit topology for very low power, reduction of leakage currents in CMOS process thanks to SOI device for example, good performance of ZigBee are mainly due to its "sleepy" (standby mode) resulting to very weak utilization of the medium protocol (MAC). Moreover, very low cost constraints lead to innovative transceiver architectures which are little greedy in silicon area while achieving good performances. The example of Zigbee transceiver, illustrated in Figure 2, uses low-IF receiver technique. It takes the advantage of many of the desirable properties of zero-IF architectures, but avoids the DC offset and 1/f noise problems. The use of a non-zero IF re-introduces the image issue. However, when there are relatively relaxed image and neighbouring channel rejection requirements they can be satisfied by carefully designed low-IF receivers. Image signal and unwanted blockers can be rejected by quadrature downconversion (complex mixing) and polyphase filtering.
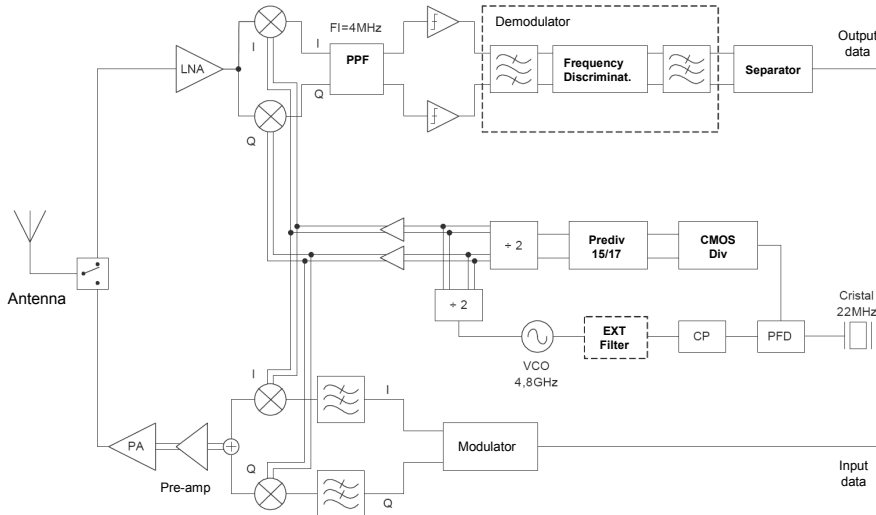
Fig. 2. Example of Zigbee transceiver proposed in (Choi et al, 2003)

In order to facilitate the complete integration of the radio section on chip with lower silicon area and lower cost, zero-IF architecture or direct-conversion receiver constitutes an efficient solution and is a good platform for multi-band multi-standard radios (e.g., 3G-WCDMA handsets and LR-WPAN). This architecture is also well adapted to analog/baseband co-design by the implementation of RF impairments compensation algorithms (e.g., DC offset, mismatch, low-frequency phase noise). An example of direct conversion receiver architecture is illustrated in Figure 3.
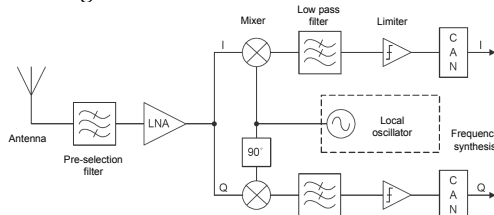
Fig. 3. Direct conversion receiver architecture

The solution adopted in the current work is a full CMOS zero-IF transceiver dedicated to very low cost and low power LR-WPAN applications and working at 2.4GHz frequency band as depicted in Figure 4. In this feature, multi-function phase locked loop (PLL) is used to synthesize 10 carrier frequencies corresponding to the transmission channels (2404 to 2488 MHz) and to modulate the data with frequency shift keying (FSK) scheme.
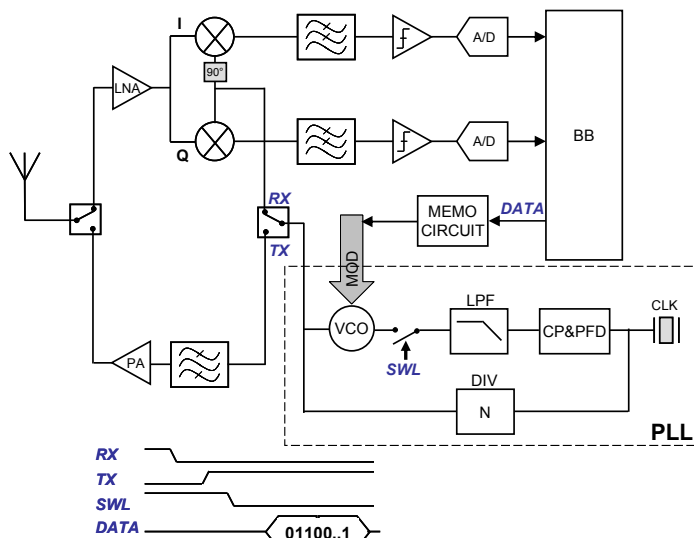
Fig. 4. IEEE 802.15.4 based transceiver architecture with direct conversion scheme.

The particularity of this solution remains in the fact that the PLL works with an open loop during the transmission mode. This provides the opportunity to completely turn off each bloc composing the PLL except the VCO and the modulation circuit allowing significant power reduction and simplify the transceiver architecture. Such simplification is possible with some modifications of the characteristics originally provided in the IEEE 802.15.4 standard including smaller channels number and therefore a larger width (10 channels of 8-MHz width), a maximum bit-rate of 125kbps and bit-error-rate (BER) of 10^{-3} instead of 50kbps and 6.10^{-5} respectively for IEEE 802.15.4.

This chapter will demonstrate the feasibility of low noise sensitivity 2.4GHz PLL for use in wireless communications in low cost LR-WPAN applications. Based on IEEE 802.15.4 specifications, this PLL is used both in a single conversion receiver as frequency synthesizer and in a direct conversion transmitter as a frequency shift keying (FSK) modulator. This multi-function low power and low cost system uses low-gain-multi-band Voltage Controlled Oscillator (VCO) which achieves a phase noise of -98dBc/Hz @ 1MHz offset while a lock time of 150µs has been obtained from the PLL loop. The circuits have been fully integrated and implemented in 130nm CMOS technology.

4. PLL design for mobile communications

In wireless communications, PLL may be used as frequency synthesizer or frequency modulation. The major challenge facing frequency synthesizer for mobile communication

devices is the need to increase their functionality in terms of operating frequency, frequency range to cover the desired operating frequency band and to accommodate process-voltage-temperature (PVT) variation, power consumption, modulation schemes while simultaneously meeting increasingly stringent linearity, phase noise and power consumption requirements at the same or lower cost. Phase locked loop (PLL) based frequency synthesizer for communication systems typically requires low phase noise and low reference spur PLLs that can be tuned over a wide range at GHz frequencies over process-voltage-temperature variations. Generally, ring oscillator ensures wide tuning range but comes with a much larger phase noise than their LC counterparts. A wide loop bandwidth is necessary to appropriately reject the high phase noise of the ring oscillator. But increasing the voltage controlled oscillator (VCO) tuning gain (K_{VCO}, in MHz/V) severally degrades the PLL phase noise and spurs performance. Meeting these conflicting requirements is the biggest challenge facing the development of future PLL modules.

4.1 Modulation circuit topology

FSK modulation has been adopted as it allows the use of power efficient, non linear RF power amplifier (McMahill & Sodini, 2002). The high tolerance to system linearity allows decreasing operating current and supply voltage. There is only a single frequency modulated carrier which is insensitive to amplifier non-linearities. Non-coherent demodulation meets the bit error rate performance requirements of this protocol and translates to simpler transceiver architectures, reducing the cost of the solution (Razavi, 1996; Razavi, 1997; Roden, 2003). Among the proposed solution in the literature concerning the frequency modulation, we can note four main methods, namely: i) sigma-delta (Σ-Δ) modulator (Huff & Draskovic, 2003; Pamarti et al, 2004), ii) two points FSK modulator (Neurauter et al, 2002), iii) Quadrature modulator and iv) two combined PLL with mixer modulator. In the very popular Σ-Δ modulation, the PLL synthesizer is directly modulated by varying the division value of the feedback divider with the output of the Σ-Δ modulator as shown in Figure 5.
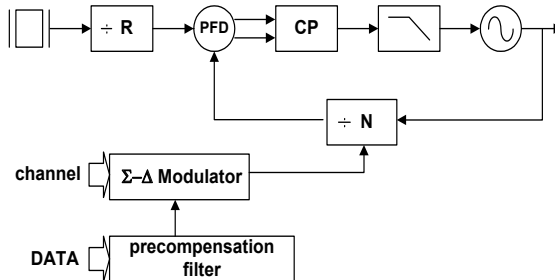
Fig. 5. Sigma-delta fractional PLL based frequency modulation

For high data rate modulation of few Mbits/s, (like DECT, CDMA2000, WCDMA), the PLL loop bandwidth (few decades of kHz) attenuates the high frequency data resulting to information lost. To overcome the limited modulation bandwidth, digital pre-emphasis filter is required (Huff & Draskovic, 2003; Pamarti et al, 2004). This operation is difficult to realize since a good matching between the analog transfer function of the PLL and the pre-emphasis digital transfer function must be ensured for proper operation. Moreover

fractional PLL is involved increasing the circuit complexity and cost. In order to bypass the loop bandwidth attenuation, more robust solution (see Figure 6) consists to apply the modulation signal at two distinct points: the low frequency signal at the Σ-Δ modulator that controls the PLL dividers while the high frequency signal is directly applied to the VCO input just after the loop filter.
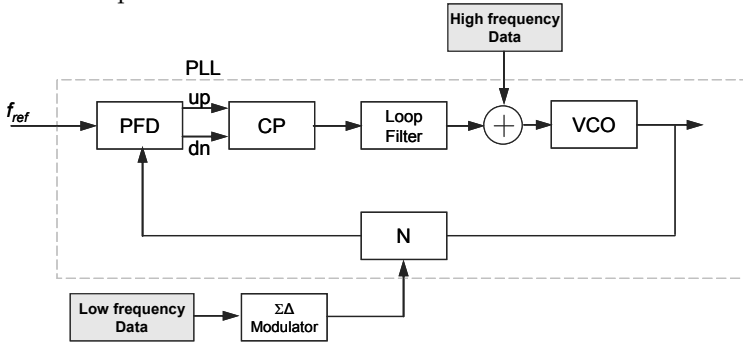
Fig. 6. GMSK two points modulation with fractional PLL

This solution requires stabilizing the VCO gain and the frequency versus temperature and process. The quadrature, or I-Q modulator, illustrated in Figure 7 is the most flexible one since any modulation type may be produced through correct choice of I(t) and Q(t) signals (McMahill & Sodini, 2002). The transmitted data sequence is processed digitally through a DSP and then converted to analog base band signal through a pair of digital-to-analog converters (DACs) to drive RF mixers whose local oscillator inputs are in quadrature. The price remains in terms of complexity and power consumption.
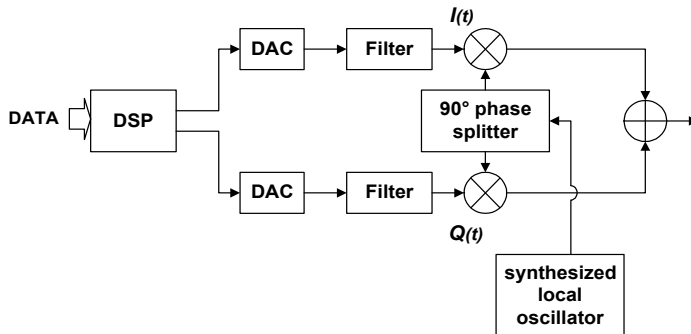
Fig. 7. Schematic of I-Q modulation method

The system illustrated in Figure 8 is a simple, low power and low cost, multi-function PLL used both in a single conversion receiver as frequency synthesizer and in a direct conversion transmitter as a frequency modulator. It can transmit or receive FSK modulated data in one of each ten 8MHz bandwidth channel and then is able to synthesize 30 frequencies from 2.404 to 2.488GHz with 2MHz step size with an open loop modulation of the VCO in the transmission mode. The modulation procedure is done in two steps: the first one is the calibration phase during which the loop is closed and the division factors are set to the first

then the second modulation frequencies. These values are injected into a memory module. The second step corresponds to the modulation phase during which the loop is opened and the data can be transmitted by directly modulating the VCO through the memory module. In terms of power consumption point of view, significant power saving is done during the transmission mode. In addition, thanks to the open loop, this modulation is not sensitive to the PLL bandwidth. However, the free running VCO is subjected to temperature, process drift and noise. The quality of the transmitted signal directly depends on its quality. During the emission phase, the entire parasitic spectrums are also amplified by the power amplifier. In order to lower the VCO sensitivity to the input noise, low gain PLL has been chosen. In fact, this solution is a simple and an efficient one for the current purpose where good signal-to-noise ratio, low phase noise and sufficiently low frequency drift are mandatory to guarantee the integrity of the emitted data. However, decreasing the conversion gain of the VCO leads to lower frequency tuning range and several VCOs are required so as to cover the entire frequency band of the transmission channels and to overcome the PVT drift. Unfortunately, this solution is not suitable for low cost, low power design requirement. In order to meet these open loop modulation constraints, new VCO topology has been proposed.
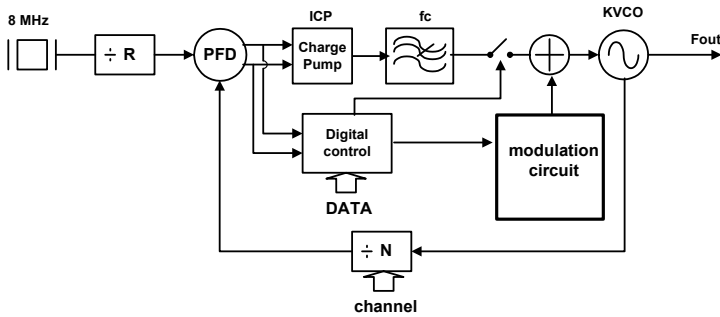
Fig. 8. Schematic of the 2.4GHz PLL and the modulator

4.2 PLL noise versus gain

The non ideality of the signal at the output of the PLL results from several design parameters. The most significant one comes from the phase noise of the VCO. But any other elements composing the PLL participates to noise degradation, the frequency divider, the noise generated by the loop filter components (thermal noise of resistors, 1/f-noise from active components), the jitter resulting from the current peak at the output of the charge-pump. Moreover, the charge-pump mismatch leads to PLL lock time degradation. Generally, PLL with a high gain generates higher noise and jitter than PLL working with lower gain. There are many solutions given in the literature in order to increase the accuracy of the synthesized frequencies. Most of them are based on the adaptive bandwidth technique (Lee & Kim, 2000; Lim et al, 2000; Vaucher, 2000) or a variant of this one. The principle is based on the modulation of the PLL bandwidth by acting on the charge-pump current together with the loop filter configuration. In fact, a closed-loop PLL can be assimilated to a low pass filter that the loop bandwidth is correlated to the PLL speed. Increasing the bandwidth can speed-up the PLL lock time, but the input noises are less filtered and degrade the spectral purity of the synthesized frequency. The adaptive

bandwidth technique is a good compromise between the PLL speed and the noise. Unfortunately this technique requires the PLL works with a closed-loop configuration and is not an efficient one if an open loop mode is required.

The solution we propose allows to solve this problematic. In fact, we propose to maintain a high charge-pump current in order to guarantee rapid lock time, while the VCO conversion gain will be decreased. The gain of individual element contributes to the overall gain in the PLL circuit where the open loop transfer function can be written such as

$$OL(s) = \frac{K\Phi \cdot K_{VCO} \cdot Z(s)}{s} \qquad (1)$$

where $K\Phi$ is the charge-pump gain, K_{VCO} is the conversion gain of the VCO, $Z(s)$ is the loop filter transfer function. Reducing the gain of VCO will lead to a reduction of the frequency tuning range and results in a less versatile PLL circuit. The gain of the charge-pump also contributes to overall gain (and jitter) of PLL circuit. However, charge-pump with lower gain will lock more slowly than a high gain charge-pump and even prevent lock from being achieved at all. Let us assume a linear transfer function of a given VCO belonging to a frequency synthesizer system such as

$$F = K_{VCO} \times (V_0 + Vnoise) = F_0 + (K_{VCO} \times Vnoise) \qquad (2)$$

where $F_0 = K_{VCO} \times V_0$ is the center frequency of the VCO, $Vnoise$ is the equivalent noise at the input of the VCO which is not filtered by the loop filter. The term $K_{VCO} \times Vnoise$ conducts to phase noise degradation of the synthesizer that directly depends on the conversion gain value. Since low gain VCO is adopted, resulting in low frequency band, more than one should be necessary in order to cover the frequency range of the system. The corresponding transfer function is illustrated in Figure 9, where the required band is ΔF with an overall VCO gain K_{VCO}. Decreasing the gain by a ratio of n reduces the noise sensitivity with the same factor, but n VCO having this low gain (K_{VCO}/n) should be required. The price remains in terms of silicon area, power consumption and circuit complexity.
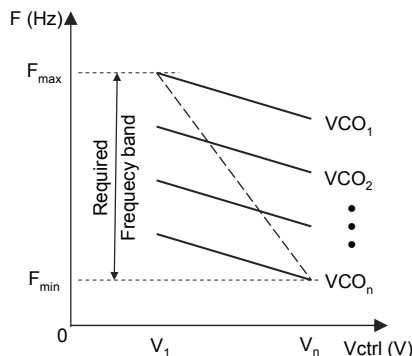
Fig. 9. Transfer function illustrating VCO conversion gain

5. Implementation of the solution

5.1 Low gain-wide-range VCO

A VCO is characterized by its operating frequency, its tuning range, its spectral purity, its power consumption, but one of the critical parameter is the phase noise. In wireless communications, phase noise requirement comes from different considerations such as interferer strength and the modulation scheme. For example, IEEE 802.11a standard (IEEE std 802.11a, 1999) uses Orthogonal Frequency Division Multiplexing (OFDM) based modulation scheme which is more sensitive to phase noise compared to single carrier modulation schemes, e.g., GMSK used in HYPERLAN Standard or FSK in LR-WPAN. Let us give an example: for a low data-rate of 125kbits/s, the current design uses FSK modulation scheme in a 4MHz channel bandwidth. Supposing –70dBm receiver sensitivity and an adjacent interferer 35dB stronger than the desired channel, the VCO phase noise needs to be lower than: PN=-35-10log(4MHz)-20 =-122dBc/Hz, assuming a predetection signal-to-noise ratio (SNR) of 20dB for a BER of 10^{-3}. This translates to a phase noise of –87dBc/Hz @ 1MHz offset. This value is reduced down to -107dBc/Hz @ 1MHz offset for the IEEE 802.11 standard (Bhattacharjee et al, 2002) where the highest data-rate equals 54Mbps using 64-QAM with OFDM in 20MHz channel bandwidth.

LC-VCO exhibits lower phase noise than its ring VCO counterpart for the same power consumption (Hajimiri, 1998; Hajimiri, 1999). Good phase noise is obtained from the high quality factor of the used inductance. Unfortunately, inductance is difficult to integrate and is area consuming; moreover its frequency control is performed thanks to a varicap diode leading to a poor frequency tuning range. In order to meet the low area requirement of the current application, ring oscillator has been chosen to implement the solution. Rather than utilizing several low gain VCOs, a solution based on single circuit that emulates the functionality of numerous VCO has been proposed. The VCO is composed by only two delay cells for an optimal integration and the corresponding schematic as given in Figure 10
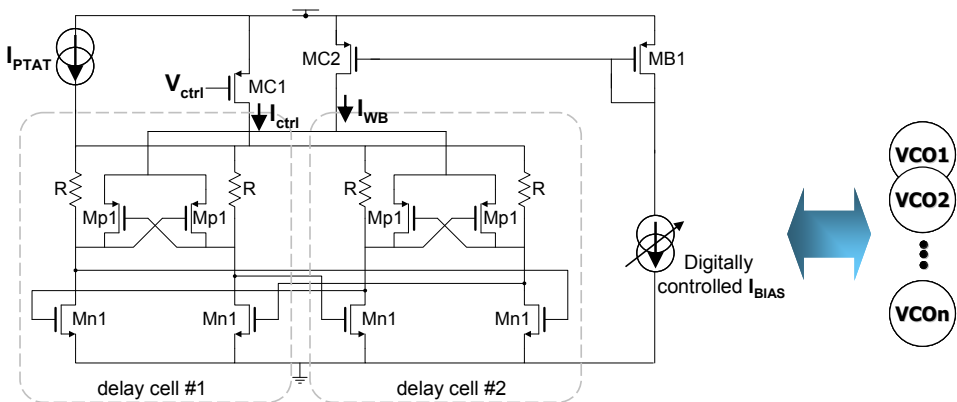
Fig. 10. Schematic of low gain wide range VCO

The circuit operation is as following. The frequency control is achieved by adjusting the transconductance (gmC) of the transistor MC1. Constant current is injected via resistor R in the differential pair Mn1. Therefore, any variation of the cross-coupled pair Mp1 transconductance (gmp1) is reported to that of Mn1 (gmn1). The bias current through the

resistors R provides an opportunity to adjust the ratio gmn1/gmp1. According to Barkhausen criteria, to maintain oscillation, the total phase of the chain has to be 360° under unity gain. At high frequencies, the current waveform may be approximated more closely by a sinusoid due to finite switching time and limited gain; as a consequence, the study can be approximated with linear model. In this condition, the oscillation frequency extracted from the delay cell related transfer function may be written as

$$f_{osc} = \frac{1}{2\pi} \sqrt{\frac{g_{mn1}^2 - \left(\dfrac{1}{R_{eq}} - g_{mp1}\right)^2}{C_{eq}^2 - C_{gdn1}^2}} \tag{3}$$

where R_{eq} is the equivalent resistance formed by R and the drain to source resistor of Mn1 and Mp1, C_{eq} represents the sum of drain to bulk and the gate to source capacitances C_{dbn1}, C_{dbp1} and C_{gsp1} of Mn1 and Mp1 respectively. As shown in (3) the output frequency directly depends on g_{mn1} and g_{mp1}. Neglecting the Early effect, the transconductance of Mn1 and Mp1 can be expressed as

$$g_{mn1} = \left[2\mu_n C_{ox} \frac{W}{L}\left(I_{bias} + I_{ctrl} + I_{WB}\right)\right]^{\frac{1}{2}} \tag{4}$$

$$g_{mp1} = \left[2\mu_p C_{ox} \frac{W}{L}\left(I_{ctrl} + I_{WB}\right)\right]^{\frac{1}{2}} \tag{5}$$

where I_{ctrl} is the current generated by the frequency tuning transistor MC1 working in ohmic regime while I_{WB} results from the programmable band selection derived from the digitally controlled current source bias, which is correlated with the channel frequency given by the digital control part of the system. Such a structure gives a possibility to emulate several low gain VCOs while covering the whole operating frequency band. The value of the VCO gain is imposed by the open loop operating mode where the effect of the voltage control drift and noise must be minimized. A VCO gain of 50MHz/V has been chosen for the current application.

5.2 Temperature compensation
During the modulation phase, the VCO is free running, and no control from the PLL loop will be applied to correct the output frequency drift from the PVT variations. Consequently, temperature compensation should be realized. The circuit topology given in Figure 10 provides the opportunity to allow double and independent frequency control. The first input is voltage controlled (gate of MC1); it corresponds to the functional input and is dedicated to transceiver parameters (action on the transconductance $g_{mn}(T)$ of Mn1 and $g_{mp}(T)$ of Mp1). The second one is current controlled (I_{PTAT} through the resistor R); it is the calibration input and is dedicated to the physical parameters (process, temperature). For the current application, it is used for temperature compensation by injecting it a temperature dependent biasing current (I_{PTAT}). In this way, the temperature control acts only on $g_{mn}(T)$, making the calibration of the compensation level extremely simple and allowing a good temperature compensation (Rahajandraibe et al, 2007).

5.3 Main loop design

The proposed PLL, illustrated in Figure 11, has been designed in the context of the current work. It is based on the charge-pump (CP) architecture.
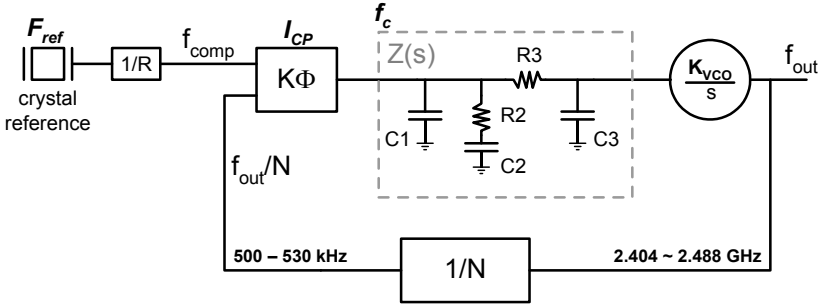

Fig. 11. PLL modeling

The 3rd order passive loop filter has been adopted in order to minimize the spurious gain, under the constraints of a constant loop bandwidth and phase margin. Its transfer function in the frequency domain can be written as

$$Z(s) = \frac{1 + s \cdot \tau_2}{s \cdot C_{tot}(1 + s \cdot \tau_1)(1 + s \cdot \tau_3)} \tag{6}$$

where $C_{tot}=C1+C2+C3$, τ is the time constant such as $\tau1=R2C2C1/C_{tot}$, $\tau2=R2C2$ and $\tau3=R3C3$. Starting from (1), the closed-loop (CL(s)) transfer function can be expressed as

$$CL(s) = \frac{N \cdot K\Phi \cdot K_{VCO} \cdot Z(s)}{s \cdot N + K\Phi \cdot K_{VCO} \cdot Z(s)} \tag{7}$$

which can be approximated by the following second order expression

$$CL_2(s) = \frac{\left(\dfrac{K\phi \cdot K_{VCO}}{N \cdot C_{tot}}\right) \cdot (1 + s \cdot N \cdot \tau_2)}{s^2 + s \cdot \left(\dfrac{K\phi \cdot K_{VCO} \cdot \tau_2}{N \cdot C_{tot}}\right)} \tag{8}$$

Defining the natural pulsation ω_n and the damping factor ξ as

$$\omega_n = \sqrt{\frac{K\Phi \cdot K_{VCO}}{N \cdot (C1 + C2 + C3)}} \qquad \text{and} \qquad \zeta = \frac{R2 \cdot C2}{2} \cdot \omega_n \tag{9}$$

and using inverse Laplace transforms, the time frequency response is obtained, from which the lock time of the PLL is derived as

$$LockTime = \frac{-\ln\left(\dfrac{tol}{f_2 - f_1} \cdot \sqrt{1 - \zeta^2}\right)}{\zeta \cdot \omega_n} \tag{10}$$

where f_2-f_1 is the frequency step, and *tol* corresponds to the maximum tolerance of the frequency at which the PLL is supposed to be locked. Each parameter of the PLL has to be chosen optimally in order to achieve low noise VCO, low power consumption, and fully integration of the loop filter capacitors. The design parameters have been calculated to meet a lock time of 150μs. That is a VCO gain of 50MHz/V, a charge-pump current ICP=3.5mA, a loop filter bandwidth of 32KHz with the following filter components: C1=10pF, C2=120pF, R2=40kΩ, C3=4pF, R3=100kΩ. The theoretical lock time computed from these values is 96μs.

6. Experimental results

In order to validate the design, test chip has been realized with 130nm CMOS technology including the entire digital and analog functional bloc composing the PLL (frequency divider, phase/frequency detector, charge-pump, loop filter, VCO, prescaler and digital blocs). A separated test chip has been implemented in the same technology for the multi-band VCO that provides the opportunity to define its intrinsic characteristics.
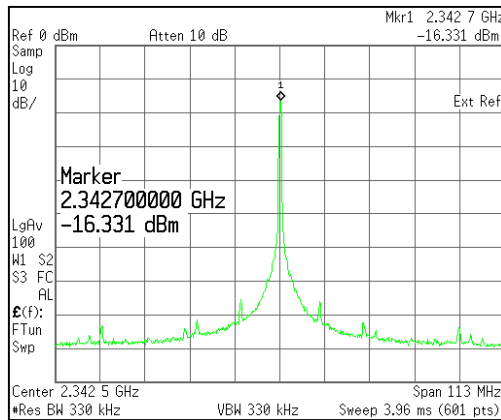
Fig. 13. Output spectrum of the low gain multiband VCO

During the modulation phase, the VCO is free running and must have sufficiently low noise and low frequency drift. The output spectrum of the free-running VCO measured with a spectrum analyzer Agilent E4446A is given in Figure 13. It exhibits an output power of -16dBm at 2.34GHz. The phase noise profile measured on the same VCO exhibits -96dBc/Hz @ 1MHz offset from the carrier. Although this value is lower than that can be achieved with an LC-VCO (-115 to -130dBc/Hz) it presents sufficient margin in view of most of the standard value required for low data rate wireless applications which, for the current application, is fixed at -87dBc/Hz @ 1MHz offset.

Figure 14 depicted the VCO output frequency tuning range as a function of the control voltage and for different configuration of the digitally programmable current source. The frequency range is shown for 10 different channels. VCO gain of 30 to 50 MHz/V is obtained for channel 1 to channel 10. Such a transfer function is equivalent to an overall gain of 90 to 100MHz/V. Lower VCO gain is obtained but the achieved performance is suitable for the current application.
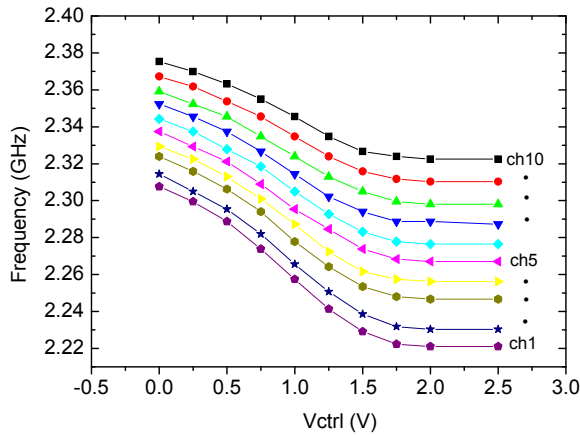
Fig. 14. Measured frequency tuning range of the VCO

The PLL has been designed and implemented in 130nm CMOS technology. A mean lock time of 100μs has been achieved when the circuit is waked up from the sleep mode to an active mode. A settling time of 36μs (see Figure 15) and 50μs are obtained between channels 1 to 2 and channels 1 to 10 respectively.
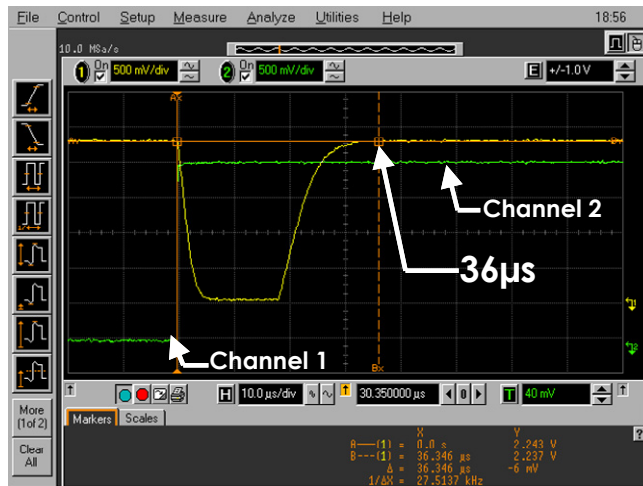

Fig. 15. Measured PLL lock time from sleep mode

A total power consumption of 34mA is obtained under a supply voltage of 2.5V. The duration of the longest pattern is 11ms. During the reception mode, the total energy consumption of the PLL equal 308μAh, however, during the transmission mode, the maximum consumption occurs only during the short locking time of the PLL, after which, the loop is opened and the data modulates directly the VCO through memory model. All the blocks are switched off except the VCO and the modulation circuit. Consequently, the system consumption is lowered down to 187μAh which represents a gain of 40%.

7. Conclusion

After arrival on the market in recent years of several wireless local area networks such as WiFi, Bluetooth, HYPERLAN and so on, news technology also appears promising a bright commercial future for both applications in public domain such as those related to home automation, and for more related field wireless communications in industrial environments: such as the ZigBee network. This WPAN network differs from its two main competitors previously cited from its simplicity of implementation and its low power consumption. ZigBee technology, coupled with the IEEE 802.15.4 standard offers simple protocol which can be declined in several versions depending on the requirement and the desired topology, for purposes of low data rate and weak of use of the medium. This article demonstrates the feasibility of high performance frequency synthesizer for this purpose. The accuracy of the output frequency is guaranteed by the low gain of the VCO without penalizing the time response (lock time) nor the frequency operating range. The design has been implemented on low cost standard CMOS technology. The proposed topology allows to realize much lower gain if it is required with a very simple calibration method.

8. References

Alliance ZigBee. *http://www.caba.org/standard/zigbee.html*.

Bhattacharjee, J., Mukherjee, D. & Laskar, J. (2002). A monolithic CMOS VCO for wireless LAN applications. *IEEE International Symposium on Circuits and Systems*, *3*, III-441 - III-444.

Chen, W.K. (2000). *The VLSI handbook.* CRC Press.

Choi, P., Park, H., Kim, S., Park, S., Nam, I., Kim, T., et al. (2003). An experimental coin-sized radio for extremely low-power WPAN (IEEE 802.15.4) application at 2.4 GHz. *IEEE Journal of Solid State Circuits*, *38* (12), 2258-2268.

Crols, J. & Steyaert, M. (1995). A single chip 900 MHz CMOS receiver front-end with a high performance low-IF topology. *IEEE Journal of Solid State Circuits*, *30* (12), 1483-1492.

Gray, P. & Meyer, R. (1995). Future directions in silicon ICs for RF personal communications., (pp. 83-90).

Hajimiri, A. & Lee, T. (1998). A general theory of phase noise in electrical oscillators. *IEEE Journal ofSolid State Circuits*, *33* (2), 179-194.

Hajimiri, A. & Lee, T. (1999). Design issues in CMOS differential LC oscillators. *IEEE Journal of Solid State Circuits*, *34* (5), 717-724.

Huff, B. & Draskovic, D. (2003, June). A fully-integrated Bluetooth synthesizer using digital pre-distortion for PLL-based GFSK modulation. *Proceedings of IEEE Radio Frequency Integrated Circuits Symposium*, 173-176.

Lee, J. & Kim, B. (2000). A Low-Noise Fast-Lock Loop with Adaptive Control. *IEEE Journal of Solid State Circuits*, *35* (8), 1137-1145.

Lim, K., Park, C.H., Kim, D.S. & Kim, B. (2000). A Low Noise Phase Locked Loop Design by Loop Bandwidth Optimisation. *IEEE Journal of Solid State Circuits*, *35* (6), 807-815.

McMahill, D. & Sodini, C. (2002). Automatic calibration of modulated frequency synthesizers. *IEEE Transactions on Circuits and Systems–II: Analog and Digital Signal Processing*, *49* (5), 301-311.

Mikkelsen, J. (1998). Evaluation of CMOS front-end receiver architectures for GSM handset applications. *IEEE Symp. Communication Systems and Digital Signal Processing* , 164-167.

Neurauter, B., Marzinger, G., Schwarz, A., & Vuketich, R. (2002). GSM 900/DCS 1800 Fractional-N Modulator with Two-Point-Modulation. *IEEE MTT-S International Microwave Symposium Digest , 1*, 425-428.

Parmarti, S., Jansson, L. & Galton, I. (2004). A Wideband 2.4-GHz Delta-Sigma Fractional-N PLL With 1-Mb/s In-Loop Modulation. *IEEE Journal of Solid State Circuits , 39* (1), 49-62.

Rahajandraibe, W., Zaïd, L., Cheynet, V. & Bas, G. (2007, Jun). Frequency Synthesizer and FSK Modulator for IEEE 802.15.4 Based Applications. *Proceedings of IEEE Radio Frequency Integrated Circuits Symposium* , 229-232.

Razavi, B. (1996). Challenges in portable RF transceiver design. *IEEE Circuits and Devices Magazine , 12*, 12-25.

Razavi, B. (1997). *RF Microelectronics.* New Jersey: Prentice Hall.

Roden, M. (2003). *Digital communication system design.* New Jersey: Prentice Hall.

ITRS: International technology roadmap for semiconductor (2007). *Radio frequency and analog/mixed-signal technologies for wireless communications*.

Vaucher, C. (2000). An Adptive PLL Tuning System Architecture Combining High Spectral Purity and Fast Settling Time. *35* (4), 490-502.

IEEE Std.802.11a (1999). *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications-High-Speed Physical Layer in the 5 GHz Band*.

Enabling Technologies for Multi-Gigabit Wireless Communications in the E-band

Val Dyadyuk, Y. Jay Guo and John D. Bunton
CSIRO ICT Centre
Australia

1. Introduction

High data rate millimeter-wave communication systems are of growing importance to the wireless industry. This can be attributed partly to an ever-increasing demand for bandwidth and scarcity of the wireless spectrum, and partly to the decreasing cost of millimetre-wave monolithic integrated circuits (MMIC) which make transmitting and receiving devices cheap to produce. Gigabit Ethernet (GbE) has become a standard protocol for the wired data transmission and usage of 10 Gigabit Ethernet (10GbE) is rapidly increasing. While known fiber optic data transfer devices can provide multi-gigabit per second data rates, infrastructure costs and deployment time can be prohibitive for some applications. Rapidly deployable, low cost wireless links can compliment the fiber networks bridging the network gaps. Multi-gigabit wireless applications include backhaul and distributed antenna systems for the 3G/4G mobile infrastructure, enterprise connectivity, remote data storage, wireless backhaul for the Wireless Local Area Networks (WLAN) and the short range wireless personal area networks (WPAN). Wide license-free spectrum around 60 Gigahertz (GHz) is allocated in most countries worldwide. While mainstream research is focused on development of multi-gigabit short range WPAN for consumer-level applications (Yong and Chong, 2007), commercial point-to-point links in the 60 GHz band with data rates up to 1.25 Giga bits per second (Gbps) are also available from several manufacturers. However, high propagation loss due to oxygen absorption in this band and regulatory requirements limit the communication range for outdoor applications. The recent allocation of the E-band spectrum (71-76 and 81-86 GHz) in USA, Europe, Russia and Australia provides an opportunity for line of sight (LOS) links with longer range and higher data rates, ideally suited for fiber replacement and backhaul applications. Current E-band commercial point-to-point wireless links[1,2,3], are limited to speeds up to 1.25 Gbps and use simple modulation techniques like amplitude shift keying (ASK) or binary phase shift keying (BPSK) with spectral efficiency below one bit per second per Hertz (bit/s/Hz).

---

[1] [Online]: http://www.bridgewave.com/products/80GHz.cfm

[2] [Online]: http://www.loeacom.com/Loea_2710_DataSheet-0508.pdf

[3] [Online]: http://www.proxim.com/downloads/products/gigalink/DS_0407_Gigalink_US.pdf

A research prototype of the E-band multi-gigabit data rate wireless communication system that uses a multi-level digital modulation has also been developed (Dyadyuk et al., 2007d). The proposed method is applicable to systems where the radio channel bandwidth is greater than the Nyquist spectral width of the associated A/D and D/A converters. Such systems can be utilized for the E-band full-duplex wireless links with a spectral efficiency scalable from 2.4 to 4.8 bit/s/Hz for 8-PSK to 64-QAM modulations to transmit 12 to 24 Gbps. This has been proven by experimental results on the 6 Gbps prototype with spectral efficiency of 2.4 bit/s/Hz. According to our knowledge this is the highest spectral efficiency achieved to date for a millimeter wave link with a demonstrated data rate above 2.5 Gbps.

While the spectrum available in the E-band allows for the multi-gigabit-per second data rates, the achievable communication range is limited by several factors, which include atmospheric attenuation and the output power attainable by semiconductor devices due to physical constraints. Currently achievable communication range of the E-band wireless networks under various propagation conditions are evaluated in this chapter using analytical estimates and experimental results. It is shown that the performance of the fixed and ad-hoc mm-wave networks for existing and emerging applications can be further improved by implementation of the spatial power combining antenna arrays.

The main challenges in the practical realization of the proposed systems, specifically the mm-wave front end integration and computationally efficient digital signal processing methods are also discussed. In this chapter we discuss enabling technologies and challenges in the commercial realization of such systems, possibilities of further improvement of fixed wireless links performance and feasibility of the development of future ad-hoc or mobile wireless networks in the E-band.

2. Multi-gigabit links for fixed terrestrial wireless networks

2.1 Spectrally efficient multi-gigabit link

The state of the art multi-gigabit wireless technology to date has been reported in our works (Dyadyuk et al., 2007a, 2007b, 2007c, and 2007d). The proposed system solution is suitable for wireless communication systems with data rates beyond 20 Gbps. We have proposed a frequency-domain multi-channel multiplexing method[4] for improved spectral efficiency, designed a 12 Gbps system in the E-band, and built a four–channel 6 Gbps concept demonstrator. With 8PSK (phase shift keying), we achieved a spectral efficiency of 2.4 bit/s/Hz. This is the highest spectral efficiency achieved to date for a millimeter wave link with a demonstrated 6 Gbps data rate. The proposed method is applicable to systems where the radio channel bandwidth is greater than the Nyquist spectral width of the associated A/D and D/A converters. As commercially available, reasonably priced analogue-to-digital (A/D) and digital-to-analogue (D/A) converters can not operate at multi-gigabit per second speeds, digital channels operating at a lower sampling speed were used. For a single carrier modulation, the proposed frequency-domain channel multiplexing technique[4] uses the root-raised-cosine digital filters (RRC) to eliminate data aliases and relaxed frequency-response requirement of analogue anti-aliasing filter. This technique allows contiguous channels to

---

[4] Bunton, J. D.; Dyadyuk, V.; Pathikulangara, J.; Abbott, D.; Murray, B.; Kendall, R. "Wireless frequency-domain multi-channel communications". Patent application WO2008067584A1, IPC H014B1/04, H04L27/26, H04B1/06 (Priority 5 Dec 2006), 12 June 2008

abut each other without the need for guard bands and makes the efficient use of wireless spectrum. While prototype system used converters with 2 Gbps sampling speed, currently available A/D and D/A can operate at the sampling speed up to 5 Gbps reducing the number of sub-channels by the factor of 2.

A simplified block diagram of the system (Dyadyuk et al., 2007d) that uses a spectrally efficient digital modulation is shown in Figure 1. The system includes a digital interface, a digital modem, an intermediate frequency (IF) module and a wideband millimeter-wave front end having transmit and receive sections and a high-directivity antenna.
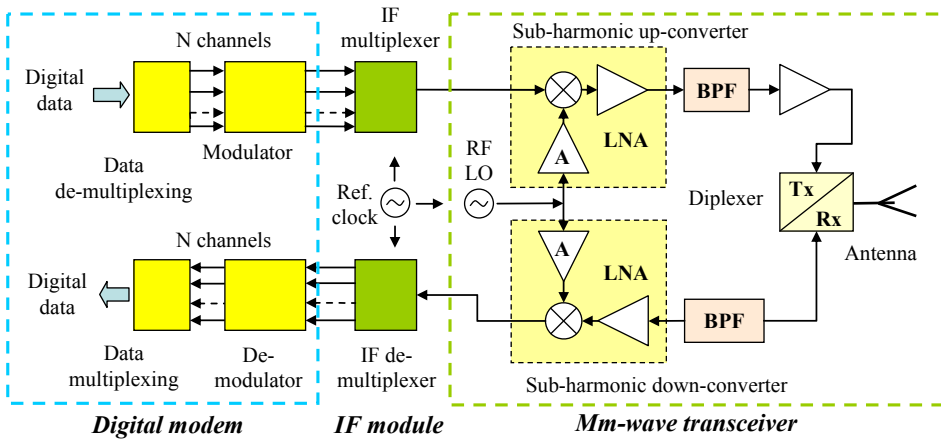
Fig. 1. Generalized block-diagram of the system

At the transmitter (Tx) input digital data stream is de-multiplexed into N digital channels (e.g. four to sixteen). At the modulator each digital channel was processed in a field-programmable gate array (FPGA) to generate the transmit symbols together with pre-compensation[5]. High speed D/As generate the analogue intermediate frequency (IF) signal for each channel in the second Nyquist band. The bands are translated in frequency and combined with the band edges abutting. No guard bands are needed because of the spectral limiting imposed by the modulation. Combined IF signal with the bandwidth equal to N •BWo is up-converted into the mm- wave carrier frequency, amplified and transmitted over a line-of-sight path.

At the receiver (Rx), received signal is down-converted from the millimeter-wave carrier frequency into IF and de-multiplexed in frequency domain into N sub-channels, then sampled by the high-speed analogue-to-digital converters (A/D) and de-coded by the FPGA that implements matched RRC filters on the *N* digital channels, these can be multiplexed into a single digital stream where required.

---

[5] Bunton, J. D.; Dyadyuk, V.; Pathikulangara, J.; Abbott, D.; Murray, B.; Kendall, R. "Wireless frequency-domain multi-channel communications". Patent application WO2008067584A1, IPC H014B1/04, H04L27/26, H04B1/06 (Priority 5 Dec 2006), 12 June 2008

For the prototype system with eight digital channels a total throughput of 12 Gbps is achieved. This can be used to implement the digital interface is either a 10 gigabit Ethernet interface together with forward error correction (FEC) on the channel. Alternatively each channel can be used to implement 1 Gbps Ethernet with FEC on the channel.

2.2 Wideband millimeter-wave transceiver

The signal to noise or interferer ratio (SNIR) required for a given BER increases with the increase in order of the multi-level digital modulations. Therefore, transceivers which can provide a high signal-to-noise ratio performance are required.

The key mm-wave transceiver in the system shown in Figure 1 uses heterodyne architectures with sub–harmonic frequency translation. Implementation of the sub-harmonic local oscillator (LO) allows a reduction in the complexity and cost of a transceiver. While a sub-harmonic mixing incurs a small penalty of a several dB in conversion gain or dynamic range, it provides a benefit of inherent suppression of both fundamental and even harmonics of the LO and down-converted LO noise.

The key element of the transceiver suitable for systems employing multi-level digital modulations is a sub-harmonically-pumped frequency converter that uses the second or fourth LO harmonic. The disadvantages compared with a fundamental LO mixer, are the slightly higher conversion loss (of about 2 dB for the 2nd harmonic), narrower bandwidth and the slightly lower conversion gain at 1dB compression level.

One convenient way to implement the architecture shown in Figure 1 entails the use of a common 39.25 GHz LO source for both receive and transmit circuits. Thus, both the 71-76 GHz and the 81-86 GHz bands can be utilized for a full-duplex communication system using the lower or upper side-band conversion in each chosen (receive or transmit) direction.

The recent progress in Si CMOS technology has largely been driven by the 60 GHz WPAN activities. Currently, the SiGe HBT and BiCMOS MMICs are the most likely candidates for high-volume 60 GHz WPANs as the reported chip sets (Cathelin et al., 2007; Floyd et al., 2007; Grass et al., (2007); Pfeiffer et al., 2008; Reynolds et al., 2007) meet current system specifications for the WPAN transceivers. This may lead to development of low-cost fully-integrated transceivers in the near future.

However, silicon chip sets suitable for the 71-76 and 81-86 GHz are not yet available. Currently, the LO driver amplifier can be built using SiGe BiCMOS, but the PA with a desired P1dB output compression of above +20 dBm are feasible only in Gallium Arsenide (GaAs) technology. Low noise SiGe amplifiers suitable for wide-band receivers have not been reported yet in the W-band.

Wide-band receive and transmit integrated modules with sub-harmonic frequency translation which were developed using a GaAs MMIC chip set have been reported in (Dyadyuk et al., 2008a, 2008b).

Figure 2 shows a photograph of the down-converter integrated into a metal housing using a traditional wire-bond approach. The LO input and the IF outputs are coaxial. The RF input uses a WR10 waveguide and an adjustable waveguide-to-microstrip transition.

The chipset includes a commercially available LNA (ALH459, Velocium, Hittite Microwave), a V-band driver amplifier (Archer and Shen, 2004) that uses a 0.15µm GaAs pHEMT process), and a sub-harmonically-pumped image-reject mixer (Dyadyuk et al., 2008a). The mixer was built using two anti-parallel pairs of 1x5 µm GaAs Schottky diodes (a standard commercial process available from United Monolithic Semiconductors).
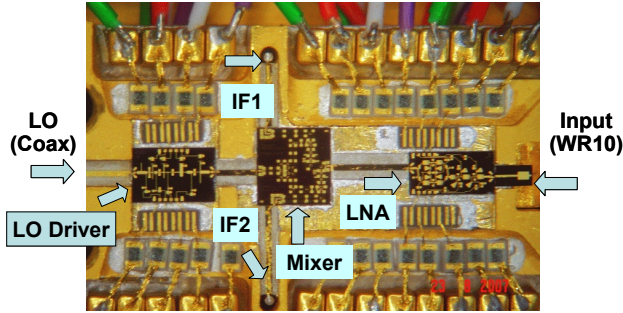
Fig. 2. Photograph of the integrated down-converter

Figure 3 shows the measured performance for an image-reject configuration using an external 90° hybrid (Krytar Model 1831). The measured image rejection was above 16 dBc. The LO power was about -9 dBm at the input to the module. The Noise Figure is estimated to be below 7.5 dB based on measured MMIC data and the insertion loss of the package inter-connects. The module exhibits extremely wideband performance with a -3 dB bandwidth greater than 9 GHz in both upper and lower side-bands.
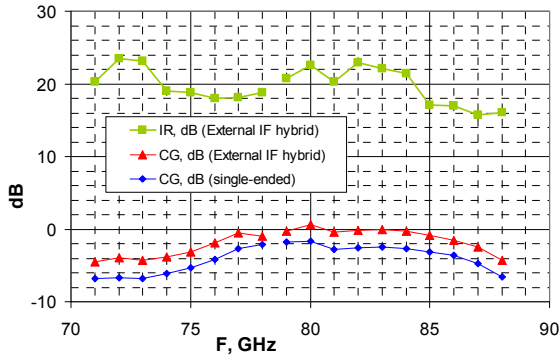
Fig. 3. Measured performance of an integrated down-converter module at the LO of 39.25 GHz in a single ended and image-reject configurations

Measured input and output P1dB compression was above -14 and -18 dBm respectively. Other experiments show that in the RF frequency range of 70 to 88 GHz the performance of the down-converter can be further optimized for a range of LO frequencies from 37 to 42 GHz resulting in a -3 dB bandwidth greater than 7 GHz in a chosen sideband.

The transmit module was integrated in a similar fashion using the same MMIC chip set with the LNA ALH459 MMIC at the output of the up-converter. The measured performance is shown in Figure 4 for a single IF port and image-reject configurations. The LO power was about -7 dBm at the input to the module. Measured 1 dB compression of the conversion gain at the IF input and RF output was above -14 and -18 dBm respectively.

An image-reject performance was measured combining the input IF ports in an external 90° hybrid (Krytar Model 1831). Measured image rejection was above 16 dBc. The measured -3

dB RF bandwidth was above 7 GHz and 5 GHz respectively in the upper and lower side-bands.
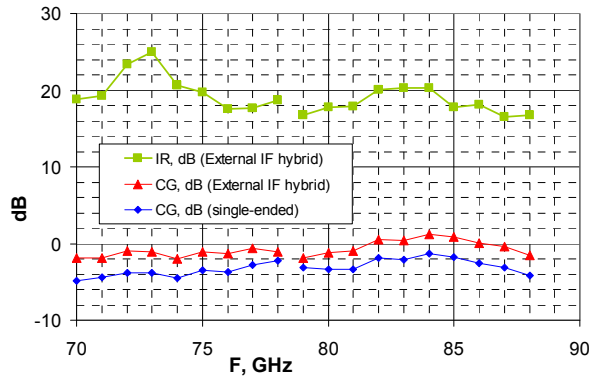
Fig. 4. Measured performance of an integrated up-converter module at the LO of 39.25 GHz in a single ended and image-reject configurations

The performance of the up-converter can be further optimized in the RF frequency band from 70 to 88 GHz for a range of LO frequencies from 37 to 42 GHz resulting in a -3 dB bandwidth of more than 7 GHz in a chosen sideband.

2.3 Frequency-domain multiplexing technique

Frequency-domain multiplexing commonly uses analogue filters that require frequency guard bands between adjacent radio channels, which is an inefficient use of the available bandwidth. The proposed method [Dyadyuk et al., 2007d] is applicable to systems where the radio channel bandwidth is greater than the Nyquist channel width of the associated A/D and D/A converters. It entails a novel frequency-domain channel multiplexing technique that combines the root-raised-cosine digital filters (RRC) to eliminate data aliases and relaxed frequency-response linear-phase analogue pass-band filters to reject only unwanted Nyquist responses without the need for guard bands.

The input binary data is de-multiplexed into N identical digital channels. A pre-compensated digital modulator is implemented in a field-programmable gate array (FPGA). Uncompensated symbols have the form of an impulse response of an RRC filter. This eliminates data aliases, and relaxes the requirements to band-pass filters (BPF) that can have up to 30% transition bands to reject unwanted Nyquist responses.

For simplicity, we describe this solution for a Return to Zero (RTZ) type of D/A converter operating at the sampling clock frequency of Fs to generate the wanted analogue signal in the second Nyquist zone. A Sync function envelope arising from the sampling by a return-to-zero (RTZ) D/A has the first zero at the double of the sampling frequency Fs. At the chosen symbol rate of Fs/4, the analogue data signal in the wanted Nyquist zone is band limited to $0.25 \cdot Fs \cdot (1+a)$, where *a* is a roll-off factor of the RRC filter, and outside this band the signal power is practically zero. The truncation of the impulse responses leads to some but low level residual power outside the wanted Nyquist zone.

The transmit sections of the system that implements the guard-band-free frequency-domain multiplexing of N high-speed digital channels (bandwidth of BWo each) into a single RF channel is shown in Figure 5.
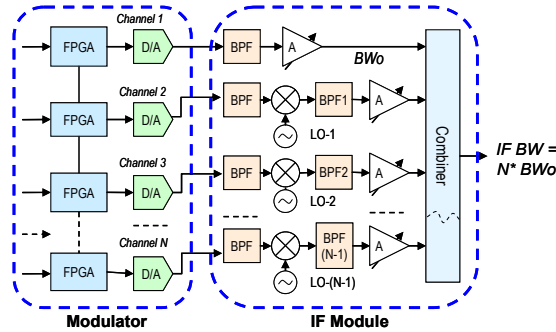

Fig. 5. The modulator and IF modules of the transmitter

The concept of combining an analogue band-pass filter (BPF) and a RRC pulse s haping filter for frequency–domain multiplexing is illustrated in Figure 6.
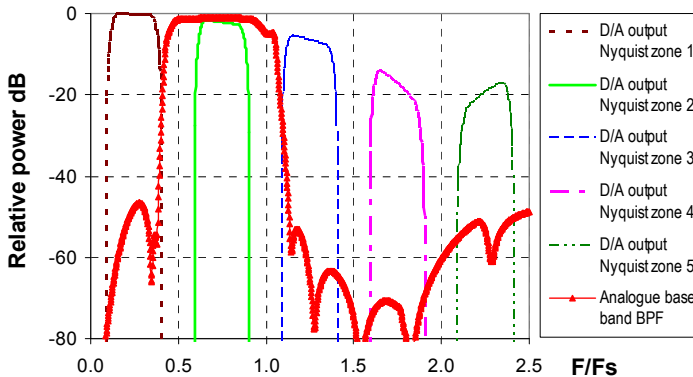

Fig. 6. First five images at the output of a RTZ D/A converter and a frequency response of a typical analogue BPF.

Figure 6 shows a D/A output in first five Nyquist zones and a typical uncompensated frequency response of an analogue BPF aligned with the second Nyquist zone. Channel 1 is directly generated by a RTZ D/A and the subsequent N-1 channels are up-converted to abut each other using frequency translation in a BWo step. Identical analogue "base band" BPF with a frequency response shown in Figure 6 is used for each digital channel at the D/A outputs. Band-pass filters BPF1 to BPF(N-1) shown in Figure 5 eliminate images arising from the frequency translation. The LO frequencies are selected to avoid unwanted mixing terms in the pass bands of neighbouring channels. This technique of using digital filters with

sharp cut-off along with the analogue band-pass filters allows contiguous channels to abut each other and allows efficient use of wireless spectrum.

A receive section that implements de-multiplexing of a receive channel into N high-speed digital channels is shown in Figure 7. The received signal is down- converted from the mm-wave carrier frequency into IF and de-multiplexed in the frequency domain into N sub-channels, then sampled by the high-speed analogue-to-digital converters (A/D), and de-coded by the FPGA that implements matched RRC filters. The de-multiplexer employs analogue filters BPF and BPF1 to BPF(N-1) identical to the filters used in the multiplexer. Data from the N digital channels can be multiplexed into a single digital stream.
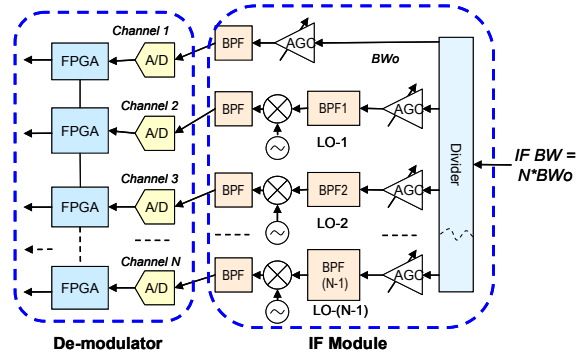
Fig. 7. The de-modulator and IF modules of the receiver

The digital modulator and demodulator are implemented in FPGAs. The FPGA logic runs at an effective sample rate Fs due to a multi-lane and parallel implementation of circuits.

The modulator stores a digital representation of the pre-compensated transmit signal for every symbol for 32 symbol periods. The symbols enter a shift register of length 32, and each of these symbols generates one set of samples from the stored representations to the output at the appropriate time. An adder chain produces the modulator waveform and the D/A converter produces the analogue IF signal. The D/A converters have a sufficient effective number of bits to accommodate this pre-compensation without degrading SINR. This novel technique is computationally efficient as no multiplications are required. Its complexity is low and grows linearly with the length of pre-compensation. The 32-symbol length modulator is sufficient to pre-compensate group delay ripple of several nanoseconds. Another innovative feature of this symbol to signal transform is that it incorporates conversion to intermediate frequency (IF), and a pre-compensated RRC filter. A chirp based channel sounding determines the pre-compensation required for the transmit symbols.

On the receive side, the FPGA digitally down converts the data from an A/D converter to the baseband quadrature (I and Q) signals. The low pass filter associated with the down converter is the RRC filter. This digital filter with a sharp cut-off rejects the out of band noise generated by frequency domain multiplexing scheme. One novel feature of this RRC filter is that it can interpolate the output sample time instant to a resolution of 1/32 of the symbol period. A bit centre tracking circuit controls the RRC sampling instant. The other blocks of the demodulator include constellation de-rotation circuits, symbol decoder, symbol insertion and deletion circuits to account for symbol rate mismatch between the transmitter and the receiver and the symbol to bits converter.

2.4 System capacity and performance

For the proposed system, the practically achievable SINR is limited by several factors. They include the LO phase noise, the limited linearity of power amplifiers (PA), inter-channel interference and the limited signal to noise and distortion ratio (SINAD) of the high speed converters over wide band pre-compensated channels. The maximum data rate is the product of the bandwidth BW and spectral efficiency E=k /(1+a), where k is the number of bits per symbol, and a is the excess bandwidth (or roll-off factor) of the root-raised-cosine (RRC) filter. The SINR required for a given BER increases with increased order of a multi-level digital modulation. Thus the SINR above 36 dB is required for k≥ 8. Whereas the phase noise of the oscillators increases with frequency, commercially available phase-locked DRO sources are suitable for the multi-level modulations. Thus, the phase noise integrated over the channel bandwidth was below – 46 dBc for the 39-42 GHz oscillators tested in the prototypes. This level is adequate for the modulations with k ≤ 8 (e.g. including 256 QAM). The measured SINAD for the commercial 2 Gsps D/A was about 50 dB for an ideal analogue channel. This was further reduced to about 40 dB for a typical physical channel. The SINAD for the A/D was measured to be about 35 dB. An approximate estimate that includes the above figures, the noise of the low-noise amplifiers, linearity of the PA and the residual inter-channel interference results in a practically attainable signal to noise or interferer ratio SINR of about 32 dB at the carrier frequency 71-86 GHz. Therefore the maximum realistic modulation order would be k ≤ 6 (e.g. 64 QAM) with E ≤ 4.8 bits/s/Hz for a typical roll-off factor of 0.25.

This leads to the conclusion that the system configuration described above can be utilized for wireless links with a spectral efficiency scalable from 2.4 to 4.8 bit/s/Hz for 8-PSK to 64-QAM modulations to transmit 12 to 24 Gbps over 5 GHz wireless bandwidth and up to 48 Gbps over 10 GHz of bandwidth.

A small-scale four-channel concept demonstrator of this system has been built using Xilinx FPGAs, Euvis model MD653 RTZ D/A converters and Atmel A/D converters operating at 2 Giga samples per second (Gsps). Four identical digital channels were multiplexed into a single 2.5 GHz wide IF signal using an optimal combination of the root-raised-cosine digital filters and linear-phase analogue filters. The base band signal bandwidth was 625 MHz at a symbol rate of 0.5 giga symbols per second and the RRC roll-of factor of 0.25. The aggregate link data rate was 6 Gbps at 2.4 Bit/s/Hz spectral efficiency for the 8PSK modulation over a 2.5 GHz width radio channel in the 81-86 GHz band.

The prototype has been installed at the 250 m long test range in Sydney, Australia. At this range a very low transmitted power of 0.25 mW was sufficient to provide link margin above 10 dB for a 99.999% annual availability at the range location at the raw bit error rate (BER) below 10^{-7}.

A separate video transmission experiment has been carried out to evaluate the link performance with a forward error corrected payload. In this experiment, sixteen video streams were aggregated into a GbE physical layer format (GMII) using multiple PCs and switches. The aggregated data of about 1.25 Gbps was transmitted over one of the digital channels using a Reed Solomon 200/216 FEC. The other three channels were used for raw BER measurements and channel sounding experiments. The video streams were generated by video cameras and DVD players. A sample of the payload received over the link is shown in Figure 8.

Fig. 8. Sample of a digital video payload received over a channel of the 6 Gbps link.

The forward-error-corrected video image was received without a loss of data or visible distortions at SINR ≥ 12 dB. This was measured with the RF path blockage of approximately 18 dB introduced by an additional RF attenuator inserted between the diplexer and antenna. The corresponding raw BER was measured on the other channels using a 1.5 Gbps Gray-coded pseudo-random 8PSK sequence. The raw BER was less than $2 \cdot 10^{-2}$.

Test results of a concept demonstrator with 6 Gbps aggregate data rate in the 81-86 GHz band and 2.4 bit/s/Hz spectral efficiency have validated the proposed system concept. In this chapter, we have chosen the above prototype as a reference point-to-point link in the predictions of the communication range of high data rate millimeter-wave communication systems.

3. Mm-wave propagation and communication range

Using the well known Frii's transmission formula, the available communication range R [km] can be determined as a root of a non-linear equation

$$P_T + G_T + G_R - 10 \cdot \log (kTB) - NF - SINR - L_0 - L_m - 92.45 - 20 \cdot \log(R) - A \cdot R - 20 \cdot \log(F) = 0 \qquad (1)$$

where P_T is transmitted power in dBW, A is the specific atmospheric attenuation in dB/km, G_T and G_R are effective gains of the receiving and transmitting antennas in dBi, k is the Boltzmann constant, T is temperature in ^oK, B is bandwidth of the receiver in Hz, NF is the noise figure of the receiver in dB, where SINR is the signal to interference and noise ratio in dB required for a certain BER by the modulation method, L_0 includes antenna pointing loss and other expected loss in dB, L_m is the minimum specified link margin in dB, and F is the frequency in GHz. The last four terms determine the LOS link free space loss.
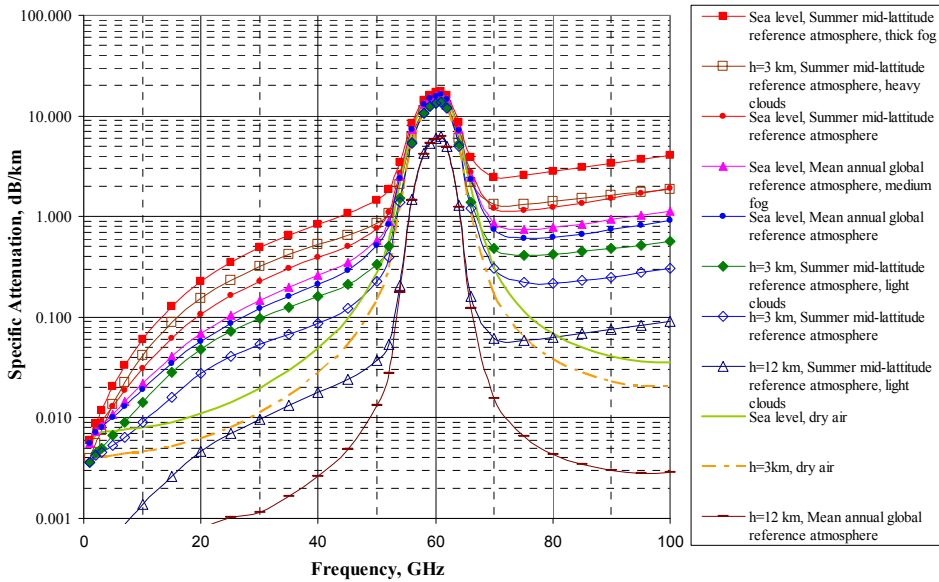
Fig. 9. Specific attenuation (in the absence of precipitation) for selected atmospheric conditions for the frequency range 10 - 100 GHz

Due to the short wavelength at mm-wave frequencies, a high gain antenna with a small physical size can be conveniently used to increase the communications range and to reduce interference with other systems. Attenuation by atmospheric gases at a specific radio frequency depends on the atmospheric conditions such as barometric pressure and temperature (both are functions of the altitude), humidity, and density of water droplets in clouds or fog. Specific attenuation A [dB/km] calculated for a horizontal path at typical atmospheric conditions at sea level and altitude h of 3 and 12 km using the ITU Recommendations[6] is given in Figure 9. Altitudes of 3 and 12 km are chosen to illustrate atmospheric attenuation for the aircraft-to-aircraft communication systems. Two standard reference atmospheres (the mean annual global reference atmosphere and the summer mid-latitude reference atmosphere) with water vapour density at sea level of 7.5 and 14.35 g/m^3 respectively are used to calculate the data given in Figure 9. Additional attenuation due to clouds and fog is estimated in accordance with the ITU Recommendation ITU-R.P.840-3 for medium fog or light clouds (visibility of the order of 300m) and thick fog or heavy clouds (visibility of the order of 50m).

It is well known that with the exception of the 60 GHz band (56-64 GHz, where radio propagation is affected by the atmospheric oxygen resonant absorption), specific attenuation increases with increasing water vapour and droplets density. In the absence of precipitation, moderate specific attenuation at the E-band (below 3 dB/km) makes this band suitable for medium and long range both terrestrial and elevated tropospheric paths. While the path loss is lower at the lower frequency, there is no current appropriate spectrum allocation at the

---

[6] ITU-R.P.676-7; ITU-R.P.840-3; ITU-R.P.510-10

frequencies below 56 GHz with the instantaneous RF bandwidth required for the multi-gigabit data rates.
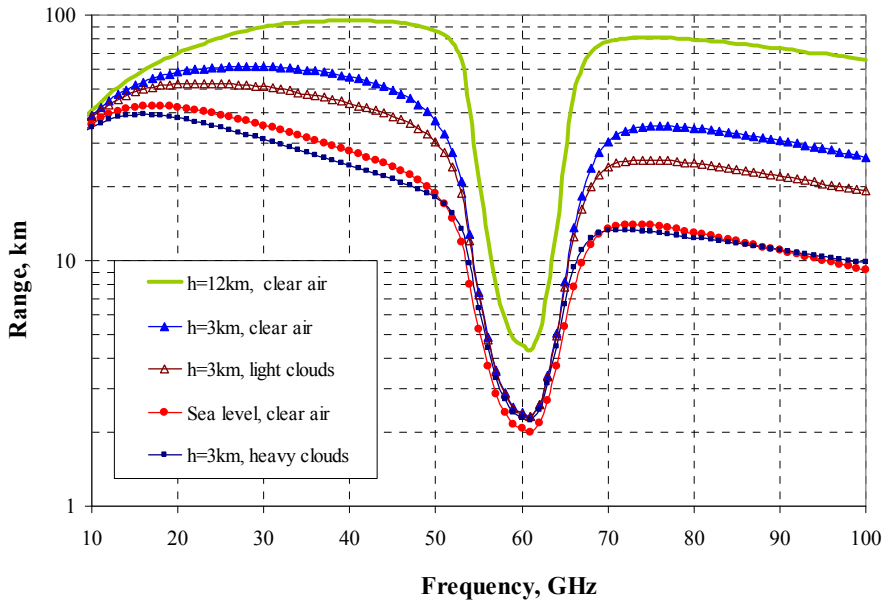
Fig. 10. Estimated communication range for a typical link with 0.36m diameter fixed beam antennas at the carrier frequency range 10 to 100 GHz

The estimated communication range for a typical configuration of the line-of-sight link equipped with identical fixed beam antennas (having a circular aperture 0.36 m diameter) is given in Figure 10 for the operating frequency from 10 to 100 GHz. For simplicity and comparison with the earlier reported results, we use a reference point-to-point link with the specification equivalent to a single 1.5 Gbps channel of the 6 Gbps prototype described above in Section 2.4 with the exception of the carrier frequency, the antenna size and transmitted power. We have assumed that a transmitter uses a single commercial power amplifier MMIC. As the data in Figure 10 was calculated for a very wide frequency range, we used linear approximation (Dyadyuk and Guo, 2009) for the output power and the receive noise figure based on the specifications of commercially available MMICs. Link margin Lm is 3dB at the bit error rate below 10^{-7} for the 8PSK. Figure 10 shows that the communication range available for the chosen link scenario does not change significantly between 10 and 100 GHz (except the 60 GHz band) at the favourable atmospheric conditions.

Hence, the frequency can be increased to take advantage of wide band operation and less interference, to achieve higher data rates over a reasonable link distance.

The main factor that limits available communication range at mm-wave frequencies is the fading due to rain. For illustrative purposes, the specific attenuation by rain Ar calculated in

accordance with the ITU-R Recommendation for vertical polarization is shown in Figure 11. The rainfall rate exceeded for a given probability of the average year for each specific location can be obtained from the ITU_R Recommendation ITU-R.P.837-5. Data Figure 11 indicates that in the upper E-band (vertical polarization) the attenuation Ar exceeds 11 dB/km and 18 dB/km respectively for heavy rainfalls of 25mm/hr and 50mm/hr.
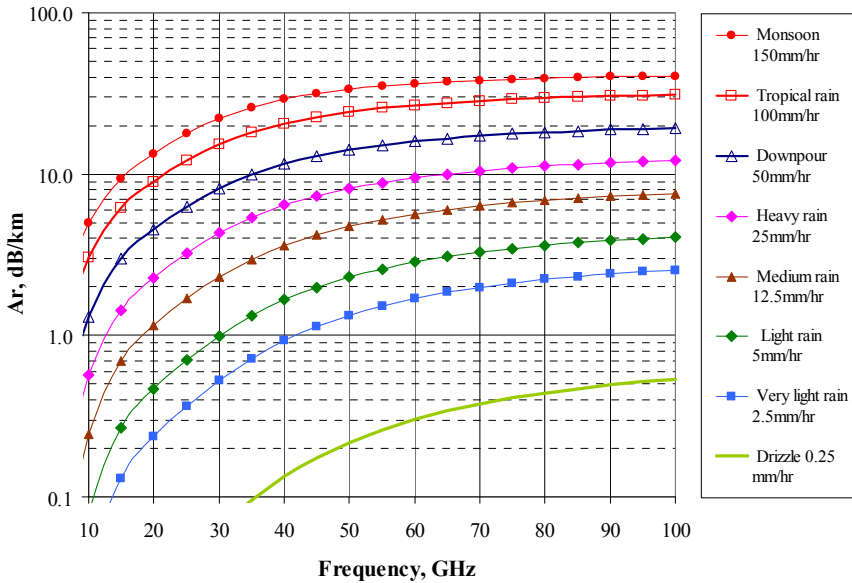
Fig. 11. Specific attenuation Ar [dB/km] due to rain (vertical polarization)

Predicted communication range for the 10 Gbps system described in Section 2 at a given rain rate is shown in Figure 12 for the carrier frequency of 83.5 GHz and 52 dBi antenna gain.
Total attenuation over a LOS path includes attenuation by atmospheric gases and rain. Multipath effects and sub-path diffraction are not modelled. The propagation model used here includes the path length reduction factor recommended by the ITU-R.P.530-12 to account for an effective rain cell size at given rain intensity and an effective path length (as rain, particularly intensive rain, is not distributed homogeneously). It is noted that the above method is recommended to 40GHz only and its use above that frequency has not been tested.
The minimum range was estimated for the BER not exceeding 10^{-7}. Adaptive modulation and appropriate forward error correction techniques can be used for operation over longer paths.
At a conservative estimate (with operating frequency selected in the upper E-band (81 – 86 GHz), and the output power of 17 dBm available from a commercial MMIC) available communication range at 10 Gbps data rate exceeds 4 km at the rain rate up to 100 mm/hr. The range can be increased using the transmit power level up to the regulatory limit of 33

dBm. The rainfall rate exceeded for a given probability of the average year can be obtained from the ITU_R Recommendation[7] for each specific location.

During more than 99.9% of the average year, very high data rate transmission over 7 – 10 km is feasible in most locations. As intensive rain events periods are very short for most locations excluding the tropical areas, required range can be maintained during these periods by using an adaptive modulation with reduced data rate as shown in Figure 12.
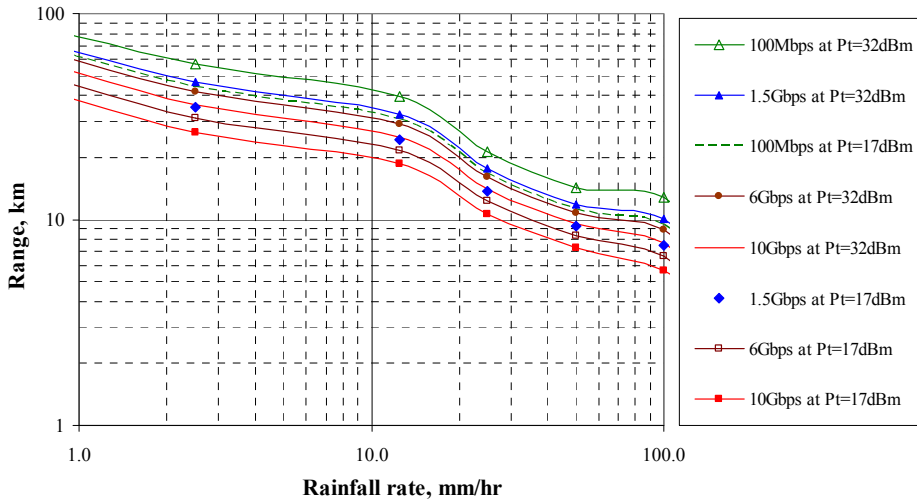

Fig. 12. Communication range of a typical E-band link versus rain fall rates. Carrier frequency is 83.5 GHz, vertical polarization, antenna gain 52 dBi.

4. Adaptive antenna arrays for future wireless communications

4.1 Spatial power combining arrays at mm-wave frequencies

With the advance in digital signal processing techniques, the adaptive antenna array is becoming an essential part of wireless communications systems (Guo, 2004; Mailoux, 2005). The use of adaptive antenna array for long range millimeter wave ad-hoc communication networks is particularly critical due to increased free space loss and reduced level of practically achievable output power. An ad-hoc or mobile network that relies on high gain antennas also requires beam scanning. The antenna beam can be steered to a desired direction with appropriate beam forming. Passive phased arrays generally suffer from losses in combining networks that are very high at the mm-wave frequencies. Active arrays with integrated power amplifier and antenna elements are effective in coherent spatial power combining increasing the total radiated power proportionally to the number of power amplifiers N. The advantages of spatial power combining are clearer at mm-wave frequencies because of the relatively low power and poor linearity of high-power amplifiers in these frequency bands. One way to overcome this problem is to use corporate power combining of multiple power amplifiers in parallel, as long as the incremental loss in the combining circuitry is less than the incremental

---

[7] ITU-R.P.837-5. Characteristics of precipitation for propagation modeling

gain of each additional power amplifier (York, 2001). As the electrical aperture and effective antenna gain is also proportional to the number of antenna array elements N, the effective isotropic radiated power (EIRP) is increased proportionally to N^2. Where the receive terminal is equipped with identical antenna array, an effective SNR increases proportionally to N^3 or more (due to reduction of the effective receiver noise dependent on the degree of the correlation). At the mm-wave frequencies, a phase-only beam steering becomes practical for this type of transmitting array since the size of a high EIRP array remains moderate. It can be shown that for a 1000-element array, the fractional bandwidth exceeds 7% at the scan angles within ±45°. This allows for a phase-only beam steering over the full 5 GHz wide RF channels available in the E-band.

We illustrate an advantage of an adaptive array over a fixed beam antenna with a simple example. With the reference link specification used in Figure 10, we replace a fixed beam antenna with a square lattice power combining array. Estimated communication range for the E-band ad-hoc link equipped with identical adaptive square lattice arrays of $N = n^2$ elements at both the transmit and receive RF terminals is shown in Figure 13 for practical scan angle from ±20° to ±45°, and typical propagation conditions. The transmitted power for each of the array elements is 15 dBm assuming that a single MMIC power amplifier with 1dB gain compression of 18 dBm is used for each array element and a -3 dB back off applied for the linearity required. A practical efficiency for the mm-wave antenna arrays is assumed of η=0.5. For clarity the carrier frequency is fixed at 73 GHz (in the lower E-band). Figure 13 shows that the medium to long communication range is achievable for the small to medium size arrays. It can be noted that the physical size of the array is very small. Thus, the aperture of the largest array shown in Figure 13 (n=64) is only a quarter of the aperture of a fixed beam antenna used in Figure 10.
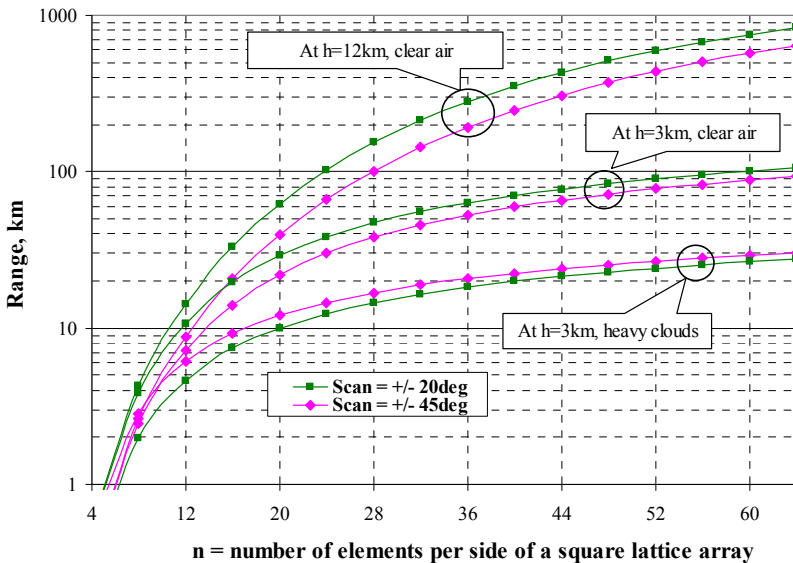
Fig. 13. Predicted communication range for a 73GHz link equipped with a square lattice power combining arrays for selected scan angles. Transmitted power is 15 dBm for each array element

For the chosen reference link parameters, a small power combining array with n = 16, N=256 elements (a linear size of 40 mm only) exhibits performance compatible with that of the link having a much larger (360mm) fixed beam antenna. A link equipped with a moderate size array, say n=36, that measures only 90 mm, is capable of the communication range beyond 100 km (at a favourable propagation conditions). Small antenna array size makes it very attractive for applications where the terminals are mounted on the mobile platforms (e.g. terrestrial and air born vehicles).

4.2 Challenges

As antenna elements must be spaced closely together (less than a half of the wavelength) to prevent grating lobes, practical realization of such antenna arrays poses a challenge due to the extremely tight space constraints at the mm-wave frequencies (about 2 mm in the E-band). The RF front end components, such as the low noise amplifier (or power amplifier), frequency converter, local oscillator (LO), as well as the intermediate frequency (IF) or baseband circuitry in the analogue signal chain should be tightly packed behind the antenna elements. With the current mm-wave integrated circuit technology, the practical implementation of such antenna arrays remains challenging. However, the recent progress in the CMOS and SiGe technology for the mm-wave applications (Cathelin et al., 2007; Floyd et al., 2007; Grass et al., 2007; Laskin et al., 2007; Pfeiffer et al., 2008; Reynolds et al., 2007) and advanced multi-chip module integration technologies (Posada et al., 2007) indicate that it becomes practical in the near future.

Although pure digitally beam forming allows the production of output signals with maximum SINR, ease of on-line calibration and generation of many antenna patterns simultaneously, it is impractical for the large wideband arrays due to two major reasons. Firstly, it is too costly since the cost of digital data processing is proportional to bandwidth and increases, at least, linearly with the number of elements. Secondly, the space constraints in the E-band make it very difficult to implement. Therefore, some degree of analogue (RF, LO or IF) beam forming is needed. This lowers the cost of digital electronics by a factor equal to the number of elements beam formed by analogue methods and also reduces the number of connections at the back of the antenna array. Thus, the area of a 4 by 4 sub-array with IF beam forming implemented in the E-band is about 100 mm² and it would provide a tight, but feasible accommodation for each the IF, LO, power and control circuits. A number of analogue sub-arrays can be controlled by a digital beam former to form a hybrid antenna array.

We have developed methods of adaptive beam forming for such hybrid arrays and built a small-scale prototype of the E-band communication system that implements an adaptive antenna array. The main functional block of the prototype is a four-channel dual-conversion receive RF module integrated with a linear end-fire steerable antenna array. Both receive and transmit modules use sub-harmonic frequency converters previously reported in (Dyadyuk et al., 2008a, 2008b). Phase and magnitude controls for each channel are implemented at IF using 6-bit phase shifters and attenuators. Both receive and transmit modules use the baseband frequency 1– 2 GHz that enables re-use of the digital modems developed earlier (Dyadyuk et al., 2007d). Bench test results of the receive and transmit modules have been reported in (Dyadyuk and Guo, 2009).

5. Conclusion

Enabling technologies for multi-gigabit spectrally efficient wireless communication systems in the E-band have been discussed. The performance of state-of-the-art E-band wireless communication system for high-capacity wireless networks has been evaluated. The analysis has been supported by experimental results on the prototypes. It has been shown that the performance of the fixed and ad-hoc mm-wave networks for existing and emerging applications can be further improved by implementation of the spatial power combining antenna arrays. The main challenges in the practical realization of the proposed systems have also been discussed.

6. Acknowledgment

7. References

Archer, J. W. and Shen, M. G. (2004). W-Band Transmitter Module Using Gallium Arsenide MMICs, *Microwave & Optical Tech. Letters,* vol. 42, no. 3, Aug. 2004, pp. 210-213, ISSN: 0895-2477

Cathelin, A.; Martineau, B.; Seller, N.; Douyere, S.; Gorisse, J.; Pruvost, S.; Raynaud, C.; Gianesello, F.; Montusclat, S.; Voinigescu, S.P.; Niknejad, A.M.; Belot, D.; Schoellkopf, J.P. (2007). Design for millimeter-wave applications in silicon technologies, *Proceedings of the 33rd European Solid State Circuits Conference,* pp. 464-471, Sep. 2007, Munich, Germany, ISSN: 1930-8833, ISBN: 978-1-4244-1125-2

Dyadyuk, V.; Stokes, L.; Sevimli, O. (2007a). A W-band multi-gigabit wireless link with high spectral efficiency, *Proceedings of* the *Intern. Joint Conf. of the TSMMW2007 and the MINT-MIS2007*, pp. 11-144, Feb. 2007, Seoul, Korea, Dongguk University, Seoul

Dyadyuk, V.; Sevimli, O.; Bunton, J. D.; Pathikulangara, J.; Stokes, L. (2007b). A 6 Gbps Millimeter Wave Wireless Link with 2.4 bit/Hz Spectral Efficiency, *Proceedings of the IEEE Intern. Microwave Symp. (IMS2007)*, pp. 471-474, June 2007, Honolulu, Hawaii, ISSN: 0149-645X, ISBN: 1-4244-0688-9

Dyadyuk, V.; Bunton, J. D.; Kendall, R.; Pathikulangara, J.; Sevimli, O.; Stokes, L. (2007c). Improved spectral efficiency for a multi-gigabit mm-wave communication system, *Proceedings of the 37th European Microwave Conf. (EuMC 2007),* pp. 810-813, Oct. 2007, Munich, Germany, ISBN: 978-2-87487-001-9

Dyadyuk, V.; Bunton, J. D.; Pathikulangara, J. et al, (2007d). A Multi-Gigabit Mm-Wave Communication System with Improved Spectral Efficiency, *IEEE Trans. on MTT, Vol. 55, Issue 12, Part 2,* Dec. 2007, pp. 2813-2821, ISSN: 0018-9480

Dyadyuk, V.; Archer, J. W.; Stokes, L. (2008a). W-Band GaAs Schottky Diode MMIC Mixers for Multi-Gigabit Wireless Communications, In: *Advances in Broadband Communication and Networks,* Agbinya, J. I. et al (Ed.), Chapt. 4, pp. 73-103, River Publishers, ISBN: 978-87-92329-00-4, Denmark

Dyadyuk, V.; Stokes, L.; Shen, M. (2008b). Integrated W-band GaAs MMIC Modules for Multi-Gigabit Wireless Communication Systems, *Proceedings of the 2008 Global Symposium on Millimeter Waves (GSMM 2008),* pp. 25-28, April 2008 Nanjing, China ISBN: 978-1-4244-1885-5

Dyadyuk, V. and Guo, Y. J. (2009). Towards Multi-Gigabit Ad-hoc Wireless Networks in the E-band, *Proceedings of the Global Symposium on Millimeter Waves (GSMM 2009),* paper 1569188415, Apr 2009, Sendai, Japan, Tohoku University, Japan

Floyd, B.; Reynolds, S.; Valdes-Garcia, A.; Gaucher, B.; Liu, D.; Beukema, T.; Natarajan, A. (2007). Silicon Technology, Circuits, Packages, and Systems for 60-100 GHz, *Proceedings of the IEEE Radio Frequency Integrated Circuits Symp. (RFIC2007), Workshop* "Millimeter-wave/Quasi–Millimeter-Wave Highly-Integrated Circuits", June 2007, Honolulu, Hawaii, ISSN: ISSN: 1529-2517, ISBN: 1-4244-0531-9

Grass, E.; Herzel, F.; Piz, M. (2007). 60 GHz SiGe-BiCMOS Radio for OFDM Transmission, *Proceedings of the 2007 IEEE Intern. Symp. On Circuits and Systems*, pp. 1979-1982, May 2007, New Orleans, ISBN: 1-4244-0920-9

Guo, Y. J. (2004). *Advances in Mobile Access Networks*, Artech House, ISBN: 1-58053-727-8, Boston, MA

Laskin, E.; Chevalier, P.; Chantre, A.; Sautreuil, B.; Voinigescu, S.P.(2007). 80/160-GHz Transceiver and 140-GHz Amplifier in SiGe Technology, *Proceedings of the IEEE Radio Frequency Integrated Circuits Symp. (RFIC2007),* pp. 153-156, June 2007, Honolulu, Hawaii, ISSN: 1529-2517, ISBN: 1-4244-0531-9

Mailloux, R. J. (2005). *Phased Array Antenna Handbook, Second Edition,* Artech House, ISBN: 1-58053-689-1, Boston, MA

Pfeiffer, U. R.; Mishra, C.; Rassel, R. M.; Pinkett, S.; Reynolds, S. K. (2008). Schottky Barrier Diode Circuits in Silicon for Future Millimeter-Wave and Terahertz Applications, *IEEE Trans. on MTT, Vol. 56, Issue 2, Feb.* 2008, pp. 364-371, ISSN: 0018-9480

Posada, G.; Carchon, G.; Soussan, P.; et al. (2007). Microstrip Thin-Film MCM-D Technology on High-Resistivity Silicon with Integrated Through-Substrate Vias, *Proceedings of the 37th European Microwave Conf. (EuMC 2007),,* pp. 1133–1136, Oct. 2007, Munich, Germany, ISBN: 978-2-87487-001-9

Reynolds, S. K.; Floyd, B. A.; Pfeiffer, U. R.; et al. (2006). A Silicon 60-GHz Receiver and Transmitter Chipset for Broadband Communications, *IEEE Journal of Solid-State Circuits*, vol. 41, 2006, pp. 2820-2831, ISSN: 0018-9200

Yong, S. K.; Chong, C. C. (2007). An Overview of Multigigabit Wireless through Millimeter Wave Technology: Potentials and Technical Challenges. *EURASIP Journal on Wireless Communications and Networking*, Vol. 2007 (2007), Article ID 78907, ISSN: 1687-1472, e-ISSN: 1687-1499

York, R.A. (2001). Some considerations for optimal efficiency and low noise in large power combiners. *IEEE Trans. on MTT,* Vol. 49, Issue 8, Aug. 2001, pp. 1477–1482, ISSN: 0018-9480

Wireless Communications at 60 GHz: A Single-Chip Solution on CMOS Technology

Chien M. Ta, Byron Wicks, Bo Yang, Yuan Mo, Ke Wang, Fan Zhang, Zongru Liu, Gordana Felic, Praveenkumar Nadagouda, Tim Walsh, Robin J. Evans, Iven Mareels, and Efstratios Skafidas

National ICT Australia (NICTA), Department of Electrical and Electronic Engineering, The University of Melbourne
Australia

1. Introduction

Consumers continuously desire improved connectivity of their electronic devices and the ability for wireless high-definition multimedia streaming, high-speed wireless file transfer, and gigabit wireless local area networks. Currently, the data rates of wireless systems utilizing the unlicensed spectrum such as Bluetooth and the IEEE 802.11 family, cannot deliver multiple-gigabit-per-second data rates due to the small bandwidth allocation. Fortunately, a 7 GHz of spectrum has been made available for unlicensed wireless communications in the 60-GHz band, from 57 to 64GHz in the USA and Canada and from 59 to 66GHz in Japan (Guo et al., 2007) to address this demand.

Ultimately for electronic devices, economic considerations decide the technology used in implementation. Systems are required to be small, low-power and low-cost. To achieve a low-cost solution, a semiconductor technology with high integration capability is needed for implementing these systems. CMOS is a standard and cost-effective process for building digital circuits. Recent advances in millimeter-wave electronics have brought a significant portion of a 60-GHz wireless system to be integrated onto a single CMOS chip. Unfortunately, compared to other more expensive processes that are capable of millimeter-wave circuit design such as SiGe and GaAs, CMOS has greater process variability, lower carrier mobility constants, and smaller device breakdown voltages. This makes 60-GHz wireless transceiver design and implementation on CMOS particularly challenging.

In this chapter, we discuss the development of a 60-GHz wireless transceiver-on-a-chip on a 130-nm CMOS technology. The challenges and solutions for the design of 60-GHz components on CMOS including radio-frequency (RF) bandpass filter (BPF), power amplifier (PA), low-noise amplifier (LNA), mixers, voltage control oscillator (VCO) are described. These components are utilized to build the world's first all-integrated 60GHz wireless transceiver on CMOS which is also presented in this chapter. The transceiver also includes a digital control interface (DCI). Experimental results are provided.

2. System architecture

In this section we outline the system architecture of the designed 60-GHz single-chip wireless transceiver. This all-integrated transceiver comprises a transmitter, a receiver, and a phase-locked loop (PLL) as shown in Fig. 1. The transmitter/receiver front-end is implemented as a homodyne architecture. The digital control interface included on the chip allows dynamical tuning of the biasing conditions of the transceiver for optimum performance.
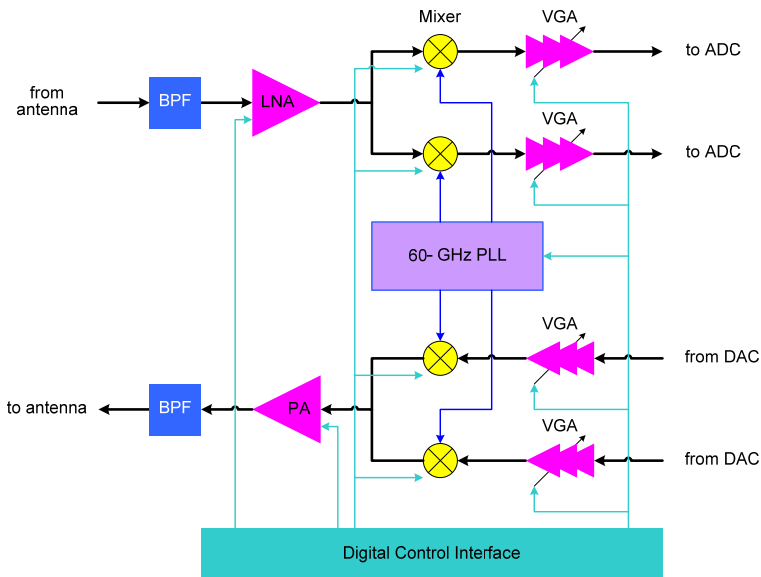

Fig. 1. Block diagram of the 60-GHz wireless transceiver

In the receiver an integrated passive BPF is employed to reject out-of-band interference to improve the receive sensitivity. The LNA amplifies the input signal while contributing a minimal amount of noise. Following the LNA are two mixers that perform frequency conversion for inphase/quadrature (I/Q) channels. These mixers are double balanced Gilbert cell mixer which offers high isolation from its local oscillator (LO) port to its RF port which is critical for homodyne transceivers (Abidi, 1995). Two variable gain amplifiers (VGAs) following the mixers have built-in DC offset cancellation loop to suppress the DC offset caused by self-mixing effect in the mixer. For the transmitter, the high-power amplifier is integrated on-chip and is optimized to drive an off-chip 50Ω-impedance transmit antenna. The BPF between the PA and the transmit antenna minimizes out of band emission in the transmitter. The local oscillation signals required for the operation of the mixers in the transmitter and receiver are provided by a PLL system. At the center of the PLL is a VCO based on a push-pull architecture. This VCO has a tuning range from 57 to 64GHz.

In the following sections, details about the design and performance of each building block of the single-chip 60-GHz transceiver on CMOS are presented.

3. Integrated passive RF filter

RF filters play an important role in radio transmitters and receivers where these filters suppress out-of-band signals generated by high-power amplifiers in the former and reject interferers to improve the sensitivity of the later. It is strongly desired to have these filters integrated on the same chip with the transceiver to reduce the overall cost and form factor of the radio. One of the biggest challenges that hinder designers from integrating RF filters on CMOS is the lossy silicon substrate. The high resistive loss induced in the silicon substrate due to electrical coupling deteriorates the quality factor of resonators implemented on CMOS which leads to higher insertion loss of filters based on these resonators (Yang et al., 2008).

The analysis and design of passive RF filter on CMOS is a challenging task due to the thick metal layers and the thin, multi-layer dielectric material on CMOS. For a thick metal trace the current distribution and the voltage potential (or E- and H-field distributions) along the top surface and those along the bottom surface are not identical. Meanwhile, the fringing coupling due to the sidewalls of the metal traces becomes significant when the metal traces grow thicker. These effects render the conventional design method which assumes thin-film metal trace on thick substrate inaccurate. In the past, where the thin-film and homogeneous, thick substrate conditions applied, theoretical and empirical design equations combined with 2D or 2.5D electromagnetic simulators are sufficient for the analysis and design of RF filters. For the thick-film, thin-substrate case of CMOS, a more rigorous approach must be taken to accurately design the filter.

In this section, the design of millimter-wave RF band-pass filters (BPFs) on CMOS technology will be presented. Along with a design methodology, methods to counter the deteriorate effects such as signal loss and coupling in millimetre-wave filters will be introduced to facilitate the realization of these filters on a CMOS technology. A BPF working on the 57-66GHz band with a compact size, a low insertion loss, and a good out-of-band rejection has been successfully implemented on the IBM 130nm CMOS technology using these techniques.
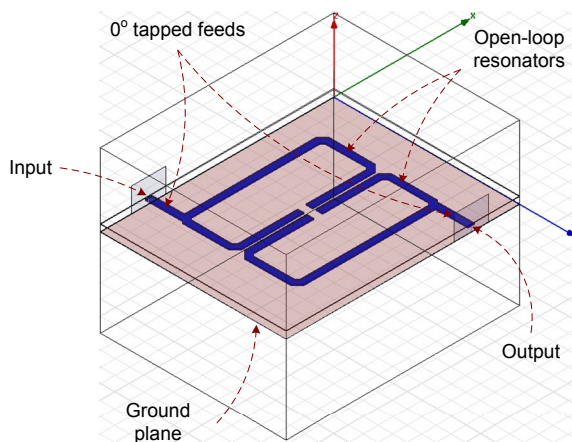
Fig. 2. HFSS model of the 60-GHz two-pole second-order open-loop resonator BPF

The designed BPF is a second-order filter based on coupled open-loop resonators. The open-loop resonators are built from half-wavelength microstrip lines. Compared to the other planar waveguide available on integrated circuit technology, the coplanar waveguide, microstrip line can be made more compact. The microstrip line also has a ground plane that shields the electro-magnetic field from coupling to the lossy silicon substrate. The microstrip structure, therefore, completely solves the problem of signal loss in the silicon substrate. The structure of the BPF as illustrated in Fig. 2 shows two rectangular resonators laid out side-by-side for electromagnetic coupling purpose. Two 0° tapped feeding structures are added to the input and output of the filter to create transmission zeros in the stopband of the filter (Lee & Tsai, 2000). These transmission zeros improve the stopband rejection of the filter. The positions at which the tapped feeding lines are connected to the resonators are optimized to ensure low ripple and low insertion loss in the passband.
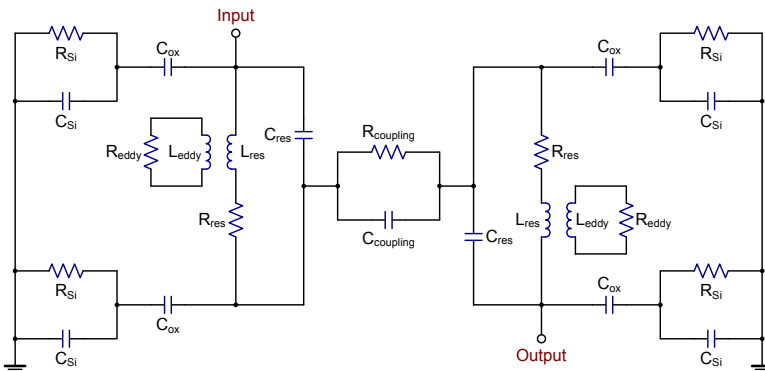
Fig. 3. A model of electromagnetic coupling between adjacent resonators on CMOS

The operation and performance of the BPF can be analyzed by investigating a lumped model of the filter. Fig. 3 illustrates the major coupling components in the two resonators of the BPF described in Fig. 2. C_{si}, and R_{si} are the capacitance and resistance, respectively, of the silicon substrate underneath the resonator. C_{ox}, is the capacitance between the metal trace of the resonator and the ground plane. C_{res} and L_{res} are the effective capacitance and inductance of the resonators. R_{res} accounts for the metal conductive loss in strips due to metal's intrinsic resistive characteristics and the skin effect that cannot be neglected at high frequencies. $C_{coupling}$ represents the proximity coupling that governs the transfer function of the filter. $R_{coupling}$ accounts for the coupling loss between two resonators. R_{eddy} represents the loss due to the eddy currents induced in the resistive substrate. The design process determines the size of the resonators and the space between them so that the desired transfer function is obtained.

The designed filter is fabricated on the IBM 130nm CMOS technology and its microphotograph is shown in Fig. 4. In order to minimize the unwanted coupling through the substrate and to reduce the induced eddy currents, the substrate was segmented into regions of high impedance. This is accomplished by implementing a high impedance substrate material between the substrate under each resonator. A high impedance bounding box is also built around the whole BPF structure. To satisfy the metal density requirements of the CMOS process, floating metal arrays are added to the layout as can be seen in Fig. 4.
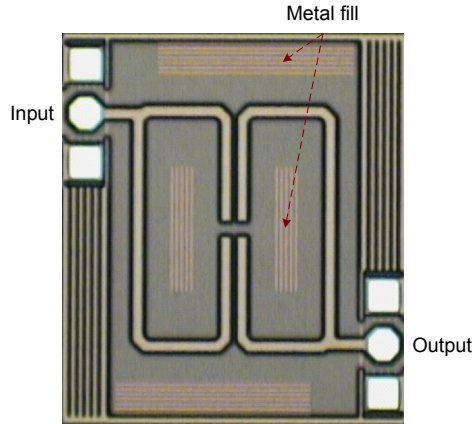
Fig. 4. Micrograph of the 60-GHz two-pole second-order open-loop resonator BPF. The footprint (excluding the testing pads) is 415μm × 503μm.
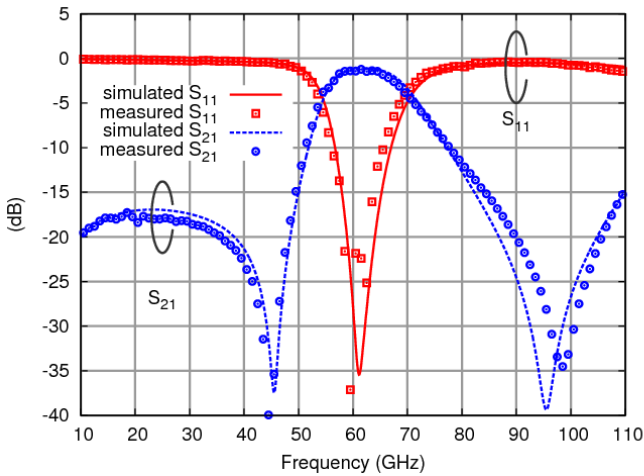

Fig. 5. Insertion loss and return loss of the 60-GHz two-pole second-order open-loop resonator BPF

HFSS, a state-of-the-art 3D full-wave electromagnetic simulator from Ansoft, Inc., is utilized to accurately estimate the performance of the BPF. Simulation and measurement results for the designed filter are shown in Fig. 5. A good match between simulation results and measurement results is achieved. The filter has a 1-dB pass-band from 57GHz to 66GHz as expected. The insertion loss is 1.5dB at mid-band frequency. The return loss is less than 9.2dB across the whole pass-band. Two transmission zeros, located at 45GHz and 98GHz, introduced by 0° tapped feed structure provide a steep roll-off in the stop-band. The side lobe in the lower stop-band is better than -18dB. This BPF has also been integrated in the 60-GHz CMOS wireless transceiver designed in this work. To the authors' knowledge, this is the first CMOS RF BPF fully integrated with a 60 GHz transceiver.

4. Low-noise amplifier

Millimeter-wave LNA design has long been dominated by compound semiconductor technologies such as GaAs and InP which have superior gain and noise performance. Only recently have silicon based technologies such as SiGe BiCMOS and CMOS, which are much cheaper than compound semiconductor technologies, advanced to a level that is capable of millimeter-wave operation. The design in (Doan et al., 2005) is the first 60-GHz amplifier on CMOS and is more a general purpose amplifier than a low-noise amplifier. The 90-nm CMOS LNA presented in (Yao et al., 2006) has a peak gain of 14.6dB at 58GHz and a minimum noise figure (NF) of 5.5dB. In general, these LNA designs are capable of providing gain for a portion of the 60-GHz band but not the entire spectrum from 57 to 66GHz, which covers all 60-GHz bands in USA, EU, and Japan, because their gain is low at the edge of the frequency band. For example, the gain of the LNA in (Yao et al., 2006) reduces to 8dB at 64GHz. Also, the NFs of these LNA are much higher at the edges of the band compared to that at the center frequency. Recent reported LNA designed on 90-nm CMOS technology (Cohen et al., 2008) achieves a very low NF of only 4.4dB. On 65-nm CMOS technology, 60-GHz LNA design was also reported to have a NF of 5.9dB and a variable gain ranging from 5dB to 15dB (Natarajan et al., 2008).
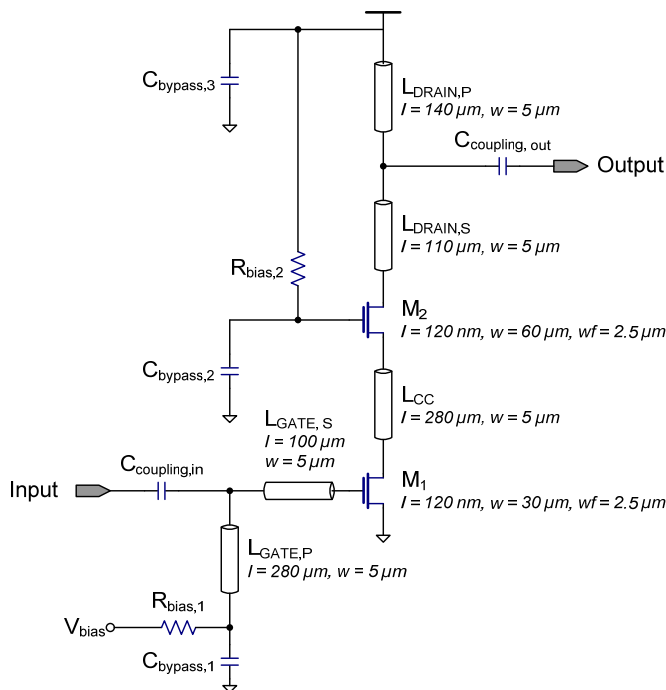
Fig. 6. Schematic of the modified cascode amplifier with intra-stage inductor

In this section the design of a 60-GHz LNA on a 130-nm CMOS process will be discussed. This LNA differs from published CMOS LNA designs in that it can operate across the entire 60-GHz band from 57 to 66GHz with less variation in gain and noise figure. The LNA which

includes four modified cascode stages was implemented and fabricated as part of the 60-GHz transceiver. Simulation results will be presented.

Millimeter-wave frequency MOSFETs are laid out as multi-finger device. Determining the right configuration of the transistor, including its gate length, gate width, and finger width, and the right biasing condition for the transistor is the crucial step to achieve the best performance of the transistor (Doan et al., 2005). For this particular CMOS technology, the optimum transistor configuration and biasing condition taken into account its gain, noise, and power consumption performance have been determined by simulation. A 130-nm gate length, 30-μm gate width, 2.5-μm finger width transistor biased at a drain current of 6mA is selected as the input transistor (M_1) of each cascode gain stage shown in Fig. 6. A major modification made to the conventional cascode stage is the adding of an intra-stage inductor L_{CC} between the input transistor M_1 and the cascode transistor M_2. This inductor helps reduce the noise figure without significantly reducing the gain of the cascode stage.
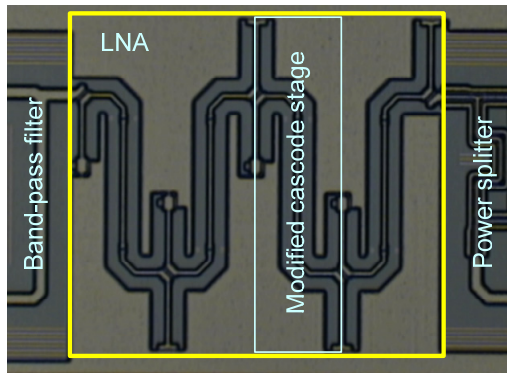
Fig. 7. Microphotograph of the four-stage LNA integrated in the 60-GHz receiver

An LNA was realized by cascading four modified cascode stages. A microphotograph of the LNA is shown in Fig. 7 as part of the receiver. Since the LNA was not taped out separately, its performance cannot be characterized experimentally. An extracted view of the LNA layout including the critical parasitic resistance and capacitance was used for the simulation of the performance of the LNA.

The performance of the LNA in the frequency band from 57GHz to 66GHz is shown in Fig. 8 (a)–(c). The maximum small-signal power gain, S_{21}, of the LNA is 17.1dB at 58GHz. From 57GHz to 65GHz, the gain of the LNA is above 14dB. The 3-dB bandwidth of the LNA is therefore 8GHz. At 66GHz, the gain is reduced to 13.1dB. The noise figure of the LNA varies from 7.2dB to 7.7dB. The input and output of the LNA match well to 50Ω with return losses, S_{11} and S_{22}, less than -12dB across the whole band. The reverse isolation, $-S_{12}$, is more than 80dB. The input third-order inter-modulation product, IIP_3, of the LNA was simulated to be 7.2dBm.

The performance of the LNA designed in this work is summarized in Table 1 along with the performance of previously published 60-GHz CMOS LNAs. These LNAs are compared by using the widely accepted Figure of Merit (FoM) defined in the International Technology Roadmap for Semiconductors (ITRS, 2007). Compared to other published CMOS LNAs

designed on 130-nm CMOS technology (Doan et al., 2005; Lo et al., 2006) the LNA in this work has higher FoM. The noise figure in this design also exhibits less variation. The comparison shows a distinct advantage of technology scaling in LNA design in which the LNAs designed in later technology, for example the 90-nm technology as reported in (Yao et al., 2007), have much higher FoM than those designed in 130-nm technology.
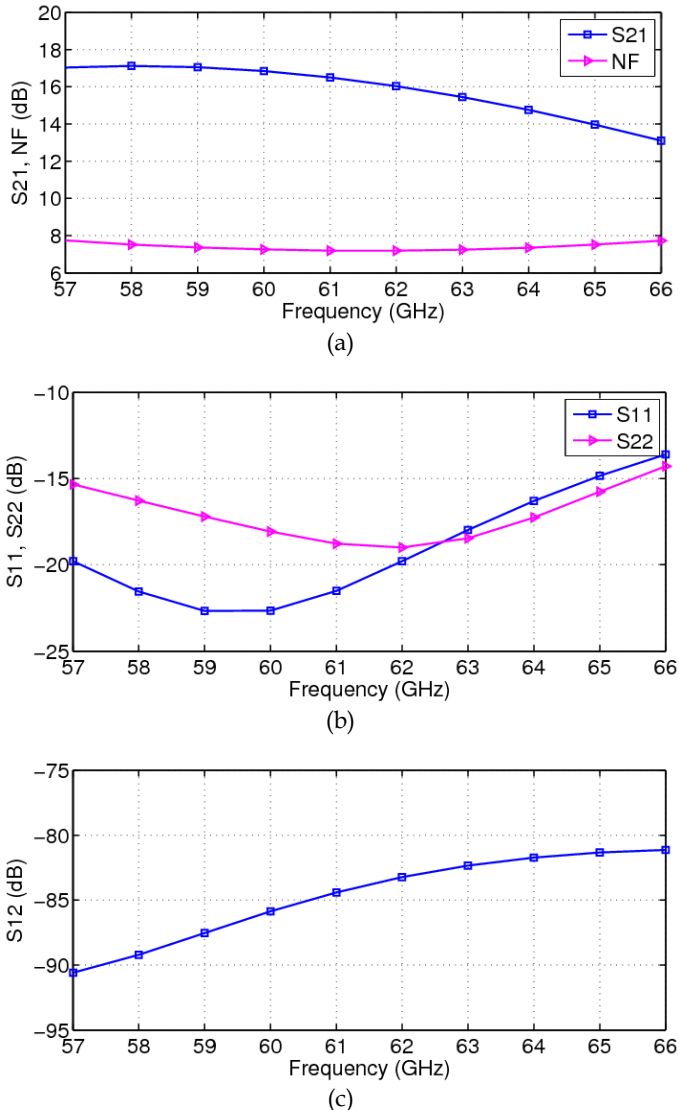

Fig. 8. Simulated performance of the LNA: (a) gain and noise figure, (b) input and output return losses, and (c) reverse isolation

Reference	CMOS tech.	Freq. (GHz)	VDD (V)	P_{DC} (mW)	NF (dB)	Gain (dB)	IIP_3 (dBm)	FoM
(Doan et al. 2005)	130-nm	51-65	1.5	54	8.8-9.7	10.4-11.9	-0.5	2.1
(Lo et al. 2006)	130-nm	51-58	2.4	72	7.1-9.5	20-24.5	-12	3.5
(Heydari et al. 2007)	90-nm	60	1.0	10.5	6.0	12.2	(n.a.)	
(Yao et al. 2007)	90-nm	58	1.5	24	4.5	14.6	-6.8	8.1
This work	*130-nm*	*57-66*	*1.5*	*36*	*7.2-7.7*	*13.1-17.1*	*-7.2*	*3.8*

Table 1. Performance comparison of the LNA in this work and previous published 60-GHz LNAs on CMOS technology

5. Power amplifier

A PA is used to amplify and efficiently deliver an RF signal it to a load which, in this design, is the transmitting antenna. At millimetre-wave frequencies the design of PAs on CMOS is challenging as the operating frequencies are a significant portion of the transistor maximum frequency of oscillation (f_{max}), thus the designer has a reduced gain available for use. As CMOS scales toward smaller feature size the breakdown voltage of the active device also reduces. This significantly limits the output power of the amplifier. The power amplifier needs to be optimised to smaller load impedances, higher current and larger device sizes to achieve a sufficient output power. Currently very few millimetre-wave power amplifiers have been published (Yao et al., 2006; Wicks et al., 2007). In addition, many implementations of high data rate systems are based on an orthogonal frequency division multiplexing (OFDM). OFDM signals are characterised by a high peak to average ratio (PAR) and require low distortion in order to minimise inter carrier interference. In order to contend with a high PAR the Doherty PA is a good architectural candidate.
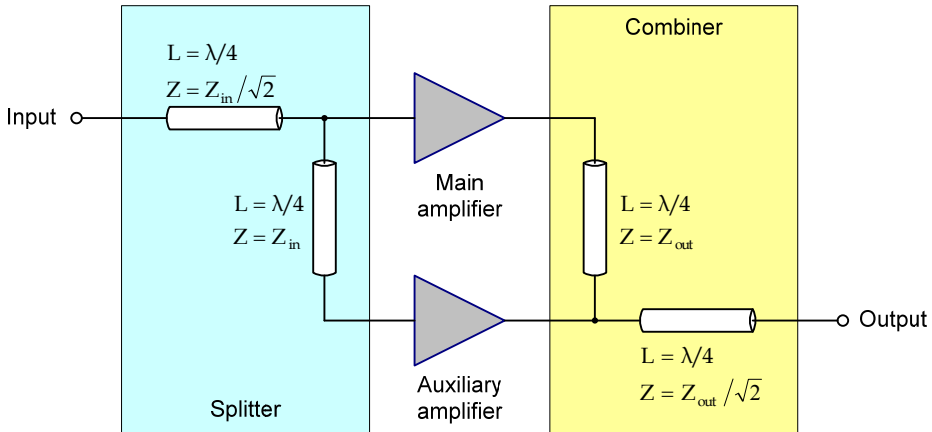
Fig. 9. Doherty power amplifier architecture

The Doherty PA employs multiple amplifiers, each contributing amplification for only a subset of the power rage and is used to boost both the power added efficiency (PAE) at low power and the 1-dB compression power (P_{1dB}) and saturation power (P_{sat}). The Doherty amplifier shown in Fig. 9 employs two amplifier cells, the main amplifier cell and the auxiliary amplifier cell. A transmission line network is used to split the input signal into two amplifiers and comprises a quarter wavelength transmission line connecting the input to the main amplifier cell with characteristic impedance of Z_0 = 50Ω, and a quarter wavelength transmission line with characteristic impedance of $Z_0 = 50/\sqrt{2}$ Ω connecting the inputs of the main amplifier cell with the auxiliary amplifier cell. An identical transmission line network is used to combine the outputs of the two amplifiers, with the output transmission line connection used to compensate for the phase shift of the splitter. Both the input and output networks are matched to 50Ω impedance.

The size of the transistors used in each main/auxiliary amplifier need to be carefully investigated. Each transistor device needs to be carefully laid out to minimise parasitic capacitance and substrate resistance. In this design each amplifier consists of five cascode stages. The cascode unit building block was used in order to increase gain of the PA by reducing the Miller capacitance, and also to improve the amplifiers stability. When biased at 200μA/μm each cascode stage possesses a maximum available gain (MAG) of 7dB at 60GHz. Micro-strip waveguides were used for impedance matching, interconnects and for biasing of each amplifying stage. These cascode stages are AC coupled to allow independent basing of the transistors for optimum operating conditions.
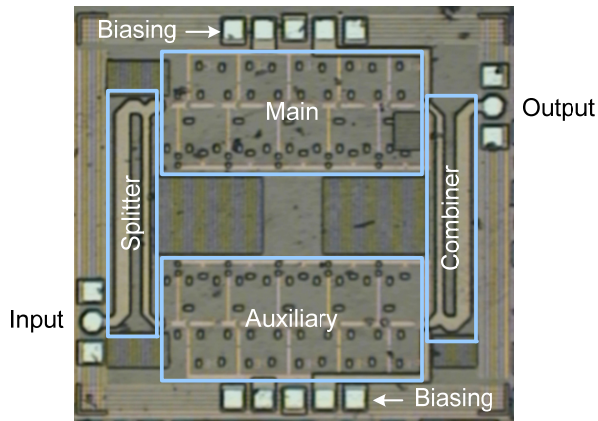
Fig. 10. Microphotograph of the 60-GHz Doherty power amplifier. The size of the PA including testing pads is 1410μm by 1310μm.

The power amplifier is fabricated on the IBM 130-nm CMOS technology and its microphotograph is shown in Fig. 10. The measured S-parameters shown in Fig. 11 reveal a peak power gain, S_{21}, of 15dB and a 3-dB bandwidth of 6GHz from 56.5GHz to 62.5GHz. The input and output return losses, S_{11} and S_{22}, are less than -10dB for the entire frequency band of interest from 57 to 66GHz. The output 1-dB compression power, P1dB, which can be derived from the high-power performance of the PA shown in Fig. 12, is 7dBm. Fig. 13 shows the current consumption in the main and auxiliary amplifier at different input power

levels. The power added efficiency (PAE) of the PA is maximized when both the main and auxiliary amplifiers are equally powered. The maximum PAE for this PA is 3%.
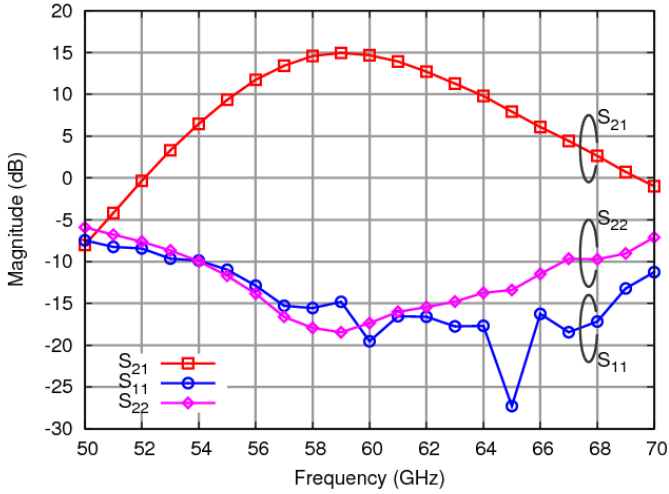

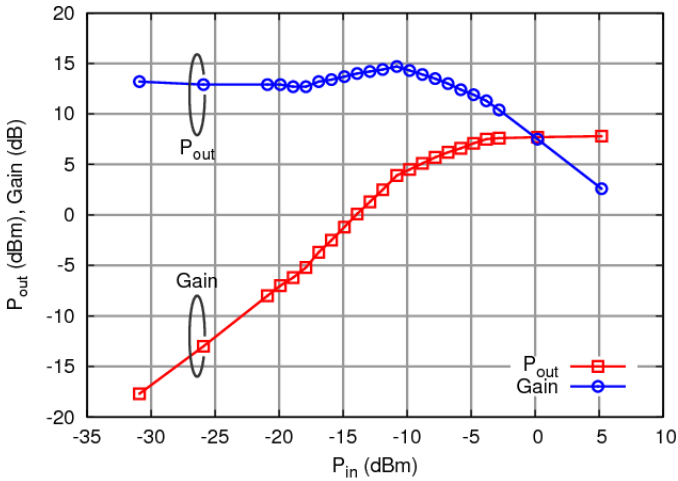Fig. 11. Measured small-signal performance of the 60-GHz Doherty power amplifier

Fig. 12. Measured output power and gain of the 60-GHz Doherty power amplifier
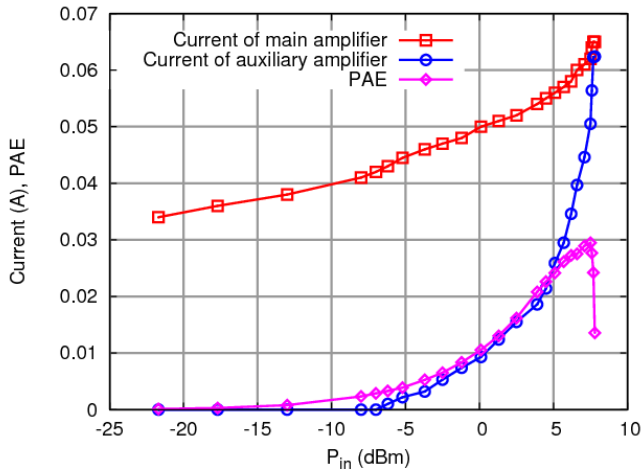
Fig. 13. Measured current consumption and power added efficiency of the Doherty PA

The International Technology Roadmap for Semiconductors (ITRS, 2007) has defined an FoM for the PA which links the output power (P_{1dB}) with the gain, PAE, and frequency as a standard to compare different PAs. Table 2 provides a comparison of this PA with other published CMOS millimeter-wave PAs in terms of this FoM.

Reference	CMOS tech.	Freq. (GHz)	Gain (dB)	P_{sat} (dBm)	P_{1dB} (dBm)	PAE (%)	Architecture	FoM
(Yao et al. 2007)	90-nm	60	5.2	9.3	6.4	7.4	3-stage cascode	7.5
(Suzuki et al. 2008)	90-nm	60	8.0	10.6	8.2	-	3-stage common source	-
	90-nm	77	9.0	6.3	4.7	-		-
(Chowdhury et al. 2008)	90-nm	60	5.6	12.3	9.0	8.8	3-stage transformer	19.5
(Wicks et al. 2008)	130-nm	77	6.0	8.1	6.3	0.5	5-stage cascode	2.1
This work	*130-nm*	*60*	*13.5*	*7.8*	*7.0*	*3.0*	*Doherty*	*15.2*

Table 2. Performance comparison of the PA in this work and published millimetre-wave PAs on CMOS technology

6. Mixer

The down-conversion mixer in the receiver is used to translate the input signal from RF to an intermediate frequency (IF) for processing by baseband circuits. An important consideration in homodyne receiver structures is the LO-to-RF isolation of the mixer. LO self-mixing (Lee, 2004), occurs when the LO signal (which is at the same frequency as the RF signal) leaks to the input of the mixer and then mixes with itself, produces a time varying DC offset which significantly degrades the receiver's performance especially in OFDM systems. In the literature, few results have been presented for CMOS mixers which are

suitable for homodyne architectures operating at the 60-GHz band (Emami et al., 2005). In this section we describe the design of a 60-GHz double balanced Gilbert cell mixer with high LO-to-RF isolation on CMOS technology.
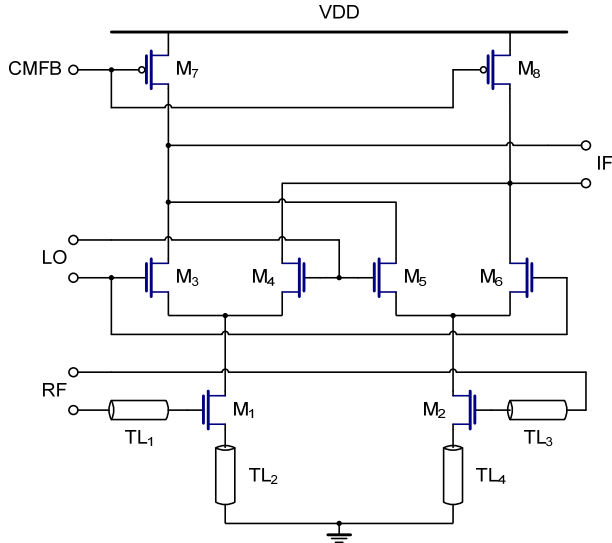

Fig. 14. Double-balanced Gilbert cell mixer

Fig. 14 shows the schematic of the double-balanced mixer where biasing circuits have been omitted for clarity. TL_2 and TL_4 are two microstrip lines serving as source degeneration inductors and TL_1 and TL_3 are used to match the RF input to 50 Ω. The input impedance looking into the transconductance stage formed by M_1 and M_2 can be shown to be equal to

$$Z_{in} = j\omega L_{TL2} + \frac{1}{j\omega C_{gs}} + \frac{g_m L_{TL2}}{C_{gs}} + j\omega L_{TL1} = j\omega L_{TL2} + \frac{1}{j\omega C_{gs}} + \omega_T L_{TL2} + j\omega L_{TL1}. \quad (1)$$

The expression above shows that by adjusting $TL_{1,2}$ we can match the input impedance to 50Ω for all different sizes of $M_{1,2}$. It also can be shown that the inductive degeneration increases linearity without raising the noise (Terrovitis, 2002). More over, by choosing the optimum number of fingers of $M_{1,2}$, the minimum noise figure, NF_{min}, can also be achieved simultaneously with input port matched.

The loads used in this Gilbert cell mixer as shown in Fig. 14 are a pair of PFET transistors. This type of load is chosen in order to achieve sufficient bandwidth and gain given the limited voltage headroom available when using a 1.2 V power supply. To get a higher output resistance for these transistors, non-minimum channel length has been used and the PFETs are biased in the strong inversion region. However, in order to drive a fixed amount of current, the longer the channel, the wider the width of a transistor is required, which may result in a large size of these PFETs. Thus, a trade-off must be made when determining the size of the PFETs. A source follower buffer, not shown in Fig. 14, is added to the output of the mixer to isolate the mixer core circuit and subsequent stages.

Because of the short wavelength of 60 GHz special considerations must be given to make the circuit as symmetrical as possible in layout to maintain balance and common mode rejection. Transmission line crossings as well as difference in path lengths are avoided when possible since these mismatches increase the imbalance and reduce the isolation between LO, RF, and IF ports. In this design, microstrip lines were used to implement the degeneration impedance, matching networks and critical interconnects that carry high-frequency signal. Micro-strip lines on silicon are typically implemented using the top- layer metal as the signal line, and the bottom-layer metal for the ground plane. The metal layers on which the signal line and the ground plane must be carefully determine so that a simple layout of the mixer can be attained without degrading the quality factor of the microstrip lines.

A 60-GHz double-balanced CMOS mixer with high LO-to-RF isolation was designed and fabricated following the design method described above. A microphotograph of the mixer is shown in Fig. 15. This mixer achieves a voltage conversion gain of better than 2dB, as shown in Fig. 16, input-referred IP3 of −8dBm and LO-to-RF isolation of greater than 36dB, as shown in Fig. 17, when driven with an LO input of 0dBm.
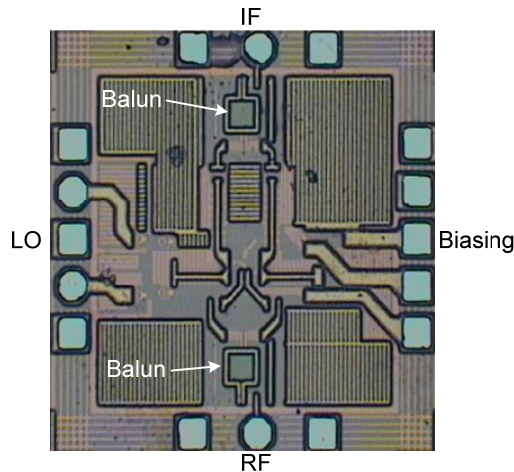
Fig. 15. Microphotograph of the 60-GHz double balance mixer on CMOS with on-chip transformer baluns for testing purpose
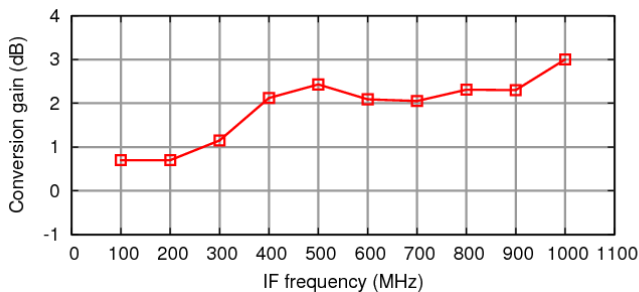
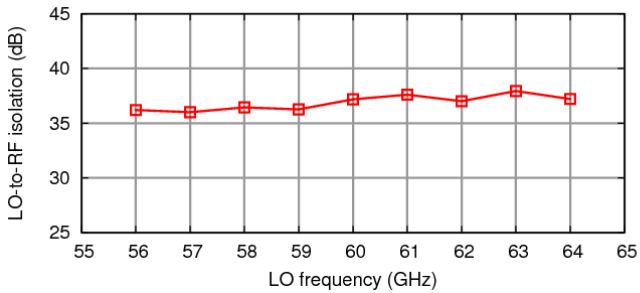Fig. 16. Conversion gain of the 60-GHz double balance mixer on CMOS

Fig. 17. LO-to-RF isolation of the 60-GHz double balance mixer on CMOS

7. Voltage controlled oscillator

The output power, tuning range and phase noise of the voltage controlled oscillator (VCO) significantly affect the performance of the transceiver. In VCO design the voltage controlled frequency of operation is achieved via voltage dependant capacitance devices such as varactors. In many cases the phase noise of these oscillators is limited by the ability to build high-quality inductors and varactors which form the LC tank that determines the frequency of the VCO. In this application the VCO is required to have: a tuning range of 9 GHz, a phase noise less than -90 dBc/Hz at 1 MHz and a sufficient output power to drive the four mixers as shown in Fig. 1. Such stringent requirements mandate a trade-off between tuning range and phase noise during the design of the VCO.

In MOS technology a varactor can be implemented by shorting the source and the drain terminals of a MOSFET together and applying a control voltage across its gate and source/drain terminals. The bias voltage governs the charge distribution in the channel and subsequently the capacitance the varactor. To achieve maximum possible tuning range with acceptably low phase noise, carefully designed inversion mode MOS varactors are employed. For this particular 130-nm CMOS technology, the length of the NMOS varactors are set equal to 260nm in order to achieve a capacitance tuning ratio of 3. An important consideration in VCO design is the gate leakage current of the varactor increases the VCO's phase noise and its parasitic capacitance reduces the oscillation frequency (Lee & Liu, 2007).

Three candidate architectures for high-frequency VCO are fundamental VCO, VCO with frequency doubler, and push-push VCO. The fundamental architecture is not very efficient in this application since it has narrow tuning range and also requires a wide band divider in the phase locked loop (PLL) which can consume significant power and space. Architectures based on frequency doubling or push-push topology are better choices because they can achieve twice the tuning range of the fundamental architecture. Another advantage of these architectures is that the varactors operate at a lower frequency and have a higher quality factor which results in reduced phase noise. Among these two architectures, the frequency doubling architecture requires additional circuits such as a doubler/multiplier and filters which can consume considerable space and power. Additionally insufficiently filtered harmonics generated by the doubler can modulate the desired output frequency of the VCO and increase the phase noise. The push-push architecture combines frequency generation and frequency doubling in one circuit. In the push-push oscillator the fundamental and odd

harmonics cancel and power is delivered to the load at even harmonics. The push-push architecture is chosen for this application.
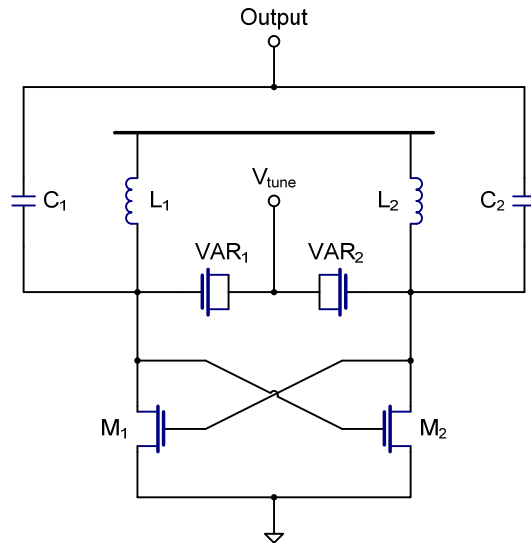

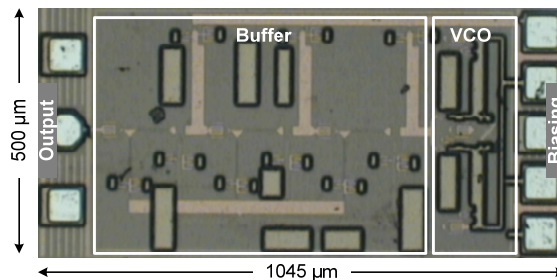Fig. 18. Circuit diagram of the push-push voltage controlled oscillator

Fig. 19. Microphotograph of the push-push voltage controlled oscillator

The differential cross-coupled LC oscillator with push-push output is shown in Fig. 18. The LC tanks composing the inductor, $L_{1,2}$, and the varactors, $VAR_{1,2}$, determines the frequency of oscillation. Frequency dependent signals at the drain of M_1 (M_2) is cross-coupled to the gate of M_2 (M_1) which creates a negative impedance $-1/g_m$ where g_m is the transconductance of $M_{1,2}$. This negative impedance is sized to exceed the losses of the LC tank to ensure sustained oscillation. Most of designs include a tail current source to set the bias current and provide high impedance which rejects noise from the power supply. However, due to the mixing effect caused by nonlinearity in $M_{1,2}$, the low frequency noise of the tail current source is up converted to the output frequency of the oscillator and degrades the phase noise of the oscillator. In this design, the current source is omitted to suppress this contribution to the phase noise. The circuit shown in Fig. 18 is implemented on standard 130nm CMOS technology. In this design the transistors $M_{1,2}$ have 50 fingers and total width

of 50µm. The NMOS varactors $VAR_{1,2}$ are implemented as a multi-finger structure to reduce gate resistance and enhance the resonator's quality factor. The inductors $L_{1,2}$ are fabricated on the top metal layer to achieve the highest quality factor possible. These inductors are realized as 100µm-long, 25µm-wide RF transmission lines and have an equivalent inductance value of 50pH. A microphotograph of the VCO is shown in Fig. 19.

The fabricated VCO has an output frequency range of 65.8GHz to 73.6GHz as shown in Fig. 20. After calibrating the cable and pads loss, the output power at 70GHz is -4dBm. The core power consumption is 32mW. The phase noise was measured by down converting the VCO's signal to an intermediate frequency of 2.5GHz. The measured phase noise is -92 dBc/Hz at 1 MHz offset (from a center frequency of 66GHz) and -107 dBc/Hz at 10 MHz offset (from a center frequency of 66GHz). Frequency variation with temperature was also measured from 0 to 70 degrees Celsius. The maximum frequency deviation is less than 200MHz in this temperature range as illustrated in Fig. 21.
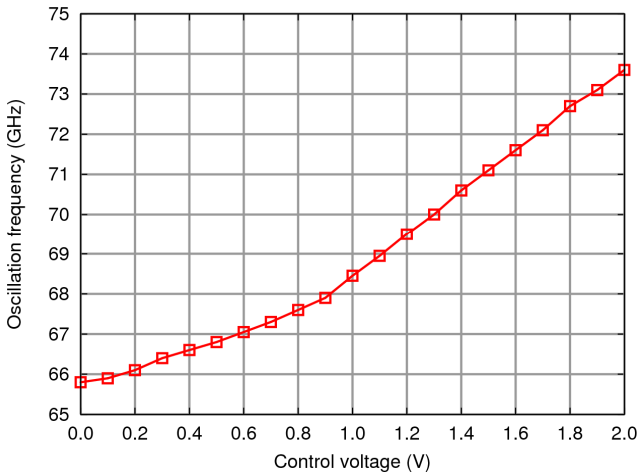
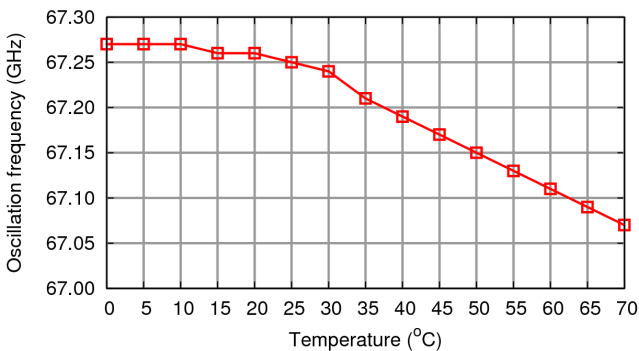Fig. 20. Output frequency versus control voltage of the VCO

Fig. 21. Frequency shift due to temperature variation

8. Biasing and control

The advancement of CMOS technology is driven by digital integrated circuits which operate faster with transistors with shorter channel length and consume less power with lower power supply voltage. These advantages of digital circuits are due to the fact that digital circuits are less sensitive to temperature, voltage, and power (PVT) variation compared to analog/RF circuits. Analog/RF circuits require special treatment during the design to reduce their sensitive to PVT variation that is significant on CMOS integrated circuits. The 60-GHz transceiver chip designed in this work adopts an on-chip Digital Control Interface (DCI) to digitally tune the behaviour of analog/RF components to remedy the performance degradation due to PVT variation thereby increasing the overall yield of the transceiver.
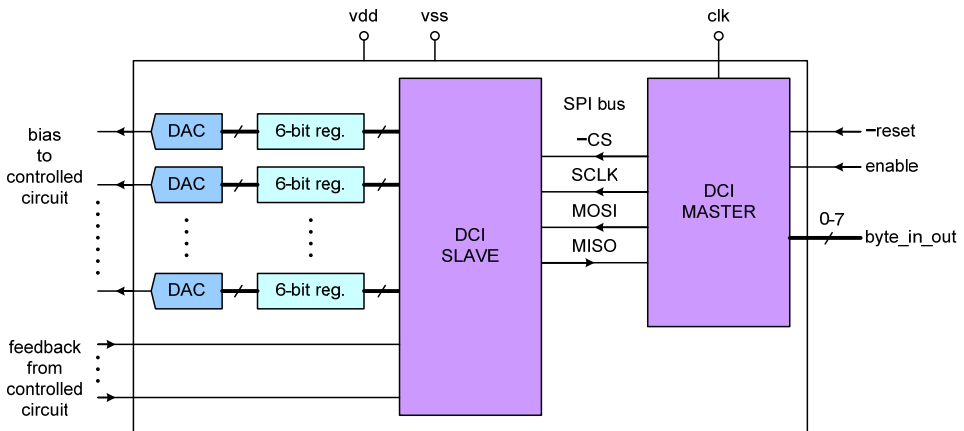

Fig. 22. Block diagram of the digital control interface

The DCI architecture is shown in Fig. 22. It comprises a DCI master and a DCI slave communicating with each other via a serial peripheral interface (SPI) bus and a bank of 6-bit registers and 6-bit digital-to-analog converters (DACs) connected to the DCI slave. For a two-chip radio solution, the DCI master resides on the digital/baseband chip while the DCI slave, register bank, and DAC bank reside on the analog/RF chip.

A tuning algorithm implemented on the digital chip will determine when a certain biasing voltage needs to be changed. The tuning algorithm will then send a request to the DCI master indicating the address of the register and the new value of the register. The DCI master passes these values to the DCI slave, via the SPI bus, which outputs the new value to the required register. The corresponding DAC translates the value stored in the register to the required analog voltage. The DCI slave can also receive feedback from analog/RF circuits and transfer this to the digital chip to assist the tuning algorithm. Real-time monitoring and tuning of the operation of the transceiver is therefore made possible with the integrated DCI.

The layout of the DCI is shown in Fig. 23. In this design, the DCI master and DCI slave are implemented together on the 60-GHz analog/RF chip.
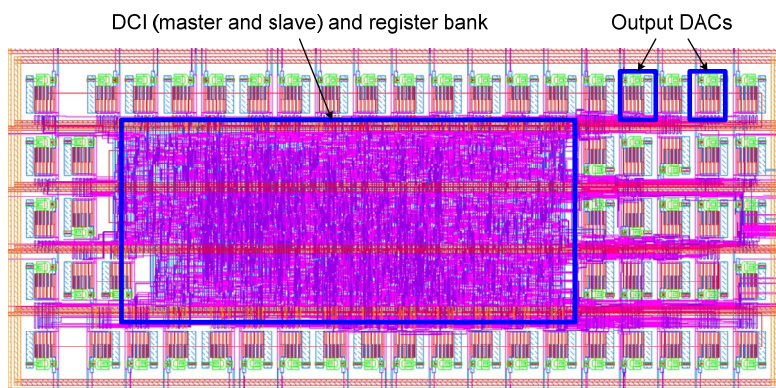
DCI (master and slave) and register bank   Output DACs

Fig. 23. Layout of the digital control interface

9. 60-GHz single-chip transceiver on CMOS

A 60-GHz single-chip CMOS transceiver was realized by integrating the circuits described above on a single silicon substrate. A microphotograph of the designed chip is shown in Fig. 24. The die measures 5mm by 5mm. Prior to this work, 60-GHz transmitters and receivers have been implemented on CMOS (Razavi, 2006; Emami et al., 2007) as well as BiCMOS (Reynolds et al., 2006). However, none of them achieved a high level of integration like this design where the transmitter and the receiver, the analog/RF circuits, the digital circuits, and the RF passive filters are all included in a single chip.
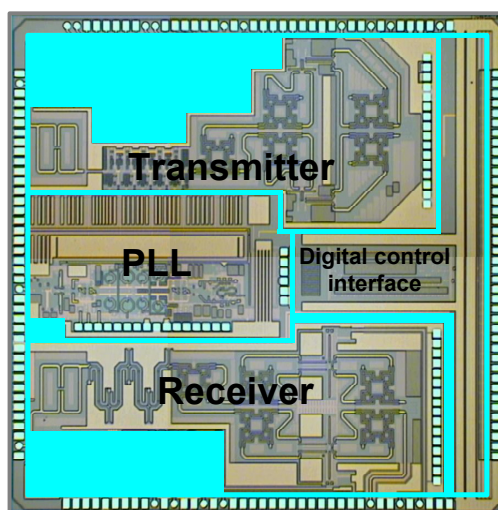

Fig. 24. The 60-GHz single-chip transceiver on 130-nm CMOS technology

The on-chip 60-GHz PLL subsystem was found not function properly even though the functionality and performance of the most challenging circuit, the 60-GHz VCO, had been

verified with measurement results as described in Section 7. An external LO signal was utilized for the purpose of demonstrating the operation of the transmitter and the receiver. The DCI functionality is satisfactory. In all measurement described below, a computer is utilized to control the DCI master. The biasing voltages for the transceiver are set by sending instructions from the computer to the DCI master via an FPGA board.
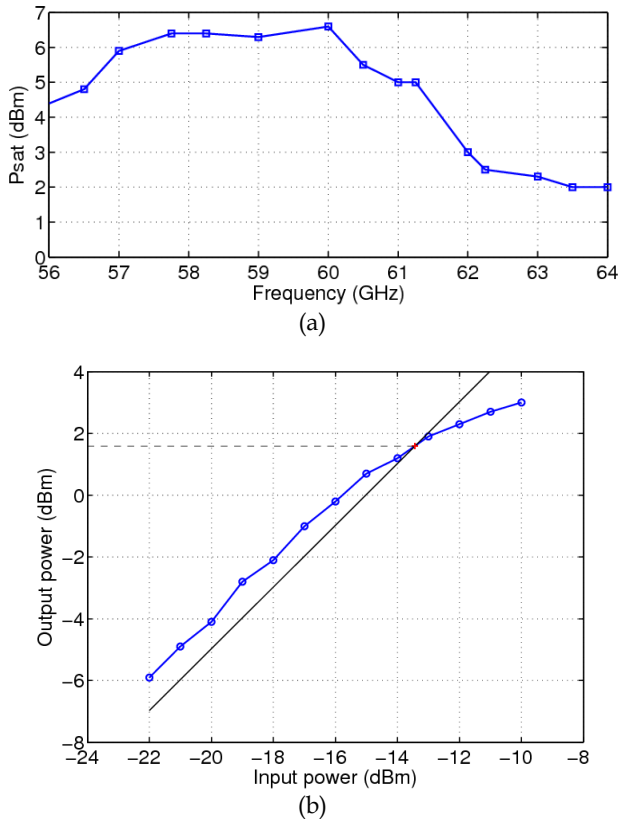

(a)

(b)

Fig. 25. Measured output power of the 60-GHz transmitter: (a) saturated output power at different output frequencies, and (b) output power versus input power at 60 GHz

The transmitter consumes a total DC power of 515mW. The transmitting capability of the transmitter is presented in Fig. 25. Fig. 25 (a) shows the saturated output power, P_{sat}, of the transmitter at different frequencies in the 56 to 64GHz band. The output power is at its peak of 6.5dBm for frequencies from 58 to 60GHz. At the high end of the spectrum, the output power is reduced to approximately 2dBm due to the degraded performance of the constituent circuits at high frequency. The output 1-dB compression power was also measured and the collected data is plotted in Fig. 25 (b). At 60 GHz, the output P_{1dB} is 1.6dBm.

The performance of the receiver including its conversion gain and linearity was measured by on-wafer probing. Noise figure measurement was not carried out due to the lack of

appropriate noise sources. The noise figure of the receiver computed from the noise figures and gains of its building blocks is 11.7dB. The receiver consumes a total power of 54mW.

The conversion gain of the receiver is presented in Fig. 26 (a) as functions of the IF frequency. A maximum conversion gain of 8.1 dB is achieved with f_{LO}=58GHz and f_{IF}=200MHz. The conversion gain of the receiver is reduced at high LO frequencies because of the reduced gain of the LNA and the down-conversion mixers at high frequencies.
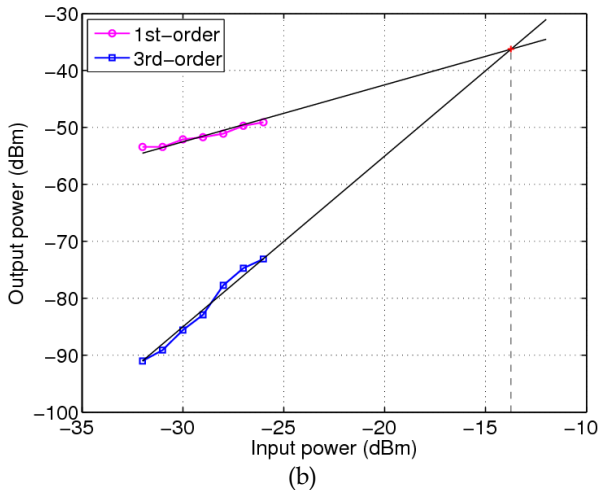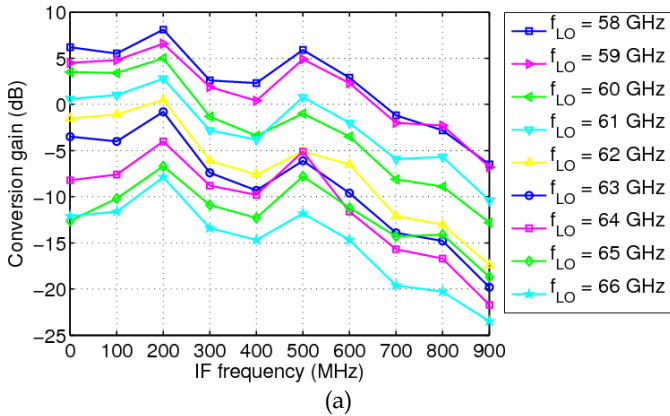

(a)

(b)

Fig. 26. Measured (a) conversion gain and (b) IIP$_3$ of the 60-GHz receiver

The linearity of the receiver, quantified by its IIP$_3$, was estimated from a two-tone test. Two Anritsu MG3690B signal generators were used to generate the two testing tones for the measurement. Due to the lack of another high power 60-GHz signal generator, the LO power was not set up properly for optimum performance of the receiver. Thus the conversion gain of the receiver was reduced substantially in this measurement. It is assumed that the output power of the fundamental tone and the third-order inter-modulation tone were reduced by the same factor so that the IIP$_3$ computed for the non-optimum operation

conditions closely tracks the IIP_3 of the receiver in its optimum operation conditions. The measured data is plotted in Fig. 26 (b). The IIP_3 of the receiver is approximately -13.74dBm.

10. Conclusion and future work

Recent advances in millimeter-wave electronics have made it possible for a complete wireless transceiver-on-a-chip system to be realized. In order to achieve a low-cost and high-integration solution CMOS is the process of choice. In this chapter we have shown the feasibility of implementing a wireless transceiver on a single chip operating in the millimeter-wave band on CMOS. The 60-GHz CMOS transceiver comprises a transmitter, a receiver, a phase-locked loop, and a digital control interface, and was implemented for the first time on a single silicon die.

The demonstration of the 60-GHz transceiver on a 130-nm CMOS process in this research, even without a working on-chip PLL, has proved the capability of CMOS technology in millimeter-wave circuit domain. However, there is still a large gap, technically and economically, that must be bridged before a truly low-cost, low-power, multi-Gbps CMOS transceiver IC can be achieved. The rest of this section discusses some future work in the process of realization such an IC.

All the circuits in this work was developed on a 130-nm CMOS technology since this was the most advanced CMOS technology characterized up to millimeter-wave frequencies at the time the research started. Owing to the fast scaling speed of CMOS technology, more advanced CMOS processes have been recently put into production by different foundries around the world. Moving the design to a more advanced technology, for example, a 65-nm CMOS technology, promises a better performance of the transceiver.

A directional, steerable phased-array antenna system is an attractive solution to overcome the high path loss at millimeter-wave frequencies and to enable transmission in non-line-of-sight conditions. The configuration of the array, i.e. one or two dimensional array, and the number of the elemental antennae, however, must be carefully determined to achieve the required link budget under certain form factor, cost, and power consumption constraints.

Since the antenna is implemented off-chip, the interface between the antenna and the CMOS transceiver across the chip boundary must be carefully studied. The interconnect between the antenna and the input/output pad on the CMOS chip, whether it is wire-bond or solder bump, must be taken into account during the design of the antenna. Further research must be carried out to understand the characteristic of these interconnects at millimeter-wave frequencies to facilitate the antenna design.

11. References

Abidi, A. (1995). Direct-conversion radio transceivers for digital communications, *IEEE Journal of Solid-State Circuits*, Vol. 30, No. 12, pp. 1399-1410, ISSN 0018-9200

Chowdhury, D.; Reynaert, P. & Niknejad, A. M. (2008). A 60GHz 1V +12.3dBm Transformer-Coupled Wideband PA in 90nm CMOS, *Digest of Technical Papers of the 2008 IEEE International Solid-State Circuits Conference*, ISBN 978-1-4244-2010-0, pp. 560-561, Feb. 2008, S³ Digital Publishing Inc., Lisbon Falls, Maine, USA

Cohen, E.; Ravid, S. & Ritter, D. (2008). An ultra low power LNA with 15dB gain and 4.4dB NF in 90nm CMOS process for 60 GHz phase array radio, *Proceedings of the 2008 IEEE Radio Frequency Integrated Circuits Symposium*, ISBN 978-1-4244-1809-1, pp. 61-64, Atlanta, Georgia, USA, June 2008, IEEE, Piscataway

Doan, C. H.; Emami, S.; Niknejad, A. M. & Brodersen, R. W. (2005). Millimeter-Wave CMOS Design, *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 1, Jan. 2005, pp. 144-155, ISSN 0018-9200

Emami, S.; Doan, C. H., Niknejad, A. M. & Brodersen, R. W. (2005). A 60-GHz down-converting CMOS single-gate mixer, *Digest of Papers of the 2005 IEEE Radio Frequency Integrated Circuits Symposium*, pp. 163-166, ISBN 0-7803-8983-2, Long Beach, California, USA, June 2005, IEEE, Piscataway

Emami, S.; Doan, C. H.; Niknejad, A. M. & Brodersen, R. W. (2007). A Highly Integrated 60GHz CMOS Front-End Receiver, *Digest of Technical Papers of the 2007 IEEE International Solid-State Circuits Conference*, pp. 190-191, ISBN 1-4244-0853-9, San Francisco, California, USA, Feb. 2007, IEEE, Piscataway

Guo, N.; Qiu, R. C.; Mo, S. S. & Takahashi, K. (2007). 60-GHz millimeter-wave radio: Principle, technology, and new results, *EURASIP Journal on Wireless Communications and Networking*, Vol. 2007, Article ID 68253, 8 pages, ISSN 1687-1472

Heydari, B.; Bohsali, M.; Adabi, E. & Niknejad, A. M. (2007). Low-power mm-wave components up to 104GHz in 90nm CMOS, *Digest of Technical Papers of the 2007 IEEE International Solid-State Circuits Conference*, ISBN 1-4244-0853-9, pp. 200–597, San Francisco, California, USA, Feb. 2007, IEEE, Piscataway

ITRS (2007). International technology roadmap for semiconductors. http://www.itrs.net/

Lee, T. H. (2004). *The Design of CMOS Radio-Frequency Integrated Circuit, 2nd Edition,* Cambridge University Press, ISBN 0-521-83539-9, Cambridge, United Kingdom

Lee, C. & Liu, S.-L. (2007). A 58-to-60.4GHz Frequency Synthesizer in 90nm CMOS, *Digest of Technical Papers of the 2007 IEEE International Solid-State Circuits Conference*, ISBN 1-4244-0852-0, pp. 196-596, San Francisco, California, USA, Feb. 2007, IEEE, Piscataway

Lee, S.-Y. & Tsai, C.-M. (2000). New Corss-Coupled Filter Design Using Improved Hairpin Resonators, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 48, No. 12, Dec. 2000, pp. 2482-2490, ISSN 0018-9480

Lo, C.-M.; Lin, C.-S. & Wang, H. (2006). A miniature V-band 3-stage cascode LNA in 0.13μm CMOS, *Digest of Technical Papers of the 2006 IEEE International Solid-State Circuits Conference*, ISBN 1-4244-0079-1, pp. 1254–1263, San Francisco, California, USA, Feb. 2006, S³ Digital Publishing Inc., Lisbon Falls, Maine, USA

Natarajan, A.; Nicolson, S.; Tsai, M.-D. & Floyd, B. (2008). A 60GHz variable-gain LNA in 65nm CMOS, *Proceedings of the 2008 IEEE Asian Solid-State Circuits Conference*, pp. 117-120, ISBN 978-1-4244-2605-8, Fukuoka, Japan, Nov. 2008, IEEE, Piscataway

Razavi, B. (2006). CMOS transceivers for the 60-GHz band, *Digest of Technical Papers of the 2006 IEEE Radio Frequency Integrated Circuits Symposium*, 4 pages, ISBN 0-7803-9572-7, San Francisco, California, USA, June 2006, IEEE, Piscataway

Reynolds, S. K.; Floyd, B. A.; Pfeiffer, U. R.; Beukema, T.; Grzyb, J.; Haymes, C.; Gaucher, B. & Soyuer, M. (2006). A Silicon 60-GHz Receiver and Transmitter Chipset for Broadband Communications, *IEEE Journal of Solid-State Circuits*, Vol. 41, No. 12, pp. 2820-2831, ISSN 0018-9200

Suzuki, T.; Kawano, Y.; Sato, M.; Hirose, T. & Joshin, K. (2008). 60 and 77GHz Power Amplifiers in Standard 90nm CMOS, *Digest of Technical Papers of the 2008 IEEE International Solid-State Circuits Conference*, ISBN 978-1-4244-2010-0, pp. 562-563, San Francisco, California, USA, Feb. 2008, S^3 Digital Publishing Inc., Lisbon Falls, Maine, USA

Terrovitis, E. (2001). Analysis and Design of Current-Commutating CMOS Mixers. *PhD dissertation*, EECS Department, University of California, Berkeley

Wicks, B.; Skafidas, E.; Evans, R. J. & Mareels, I. M. (2007). A 46.7-46.9 GHz CMOS MMIC Power Amplifier for Automotive Applications, *Proceedings of the Fourth IASTED International Conference on Antennas, Radar and Wave Propagation*, ISBN 978-0-88986-661-4, pp. 80-84, Montreal, Quebec, Canada, May-June 2007, ACTA Press, Anaheim, California, USA

Wicks, B. N.; Skafidas, E.; Evans, R. J. & Mareels, I. (2008). A 75–95 GHz Wideband CMOS Power Amplifier, *Proceedings of the 2008 European Microwave Integrated Circuits Conference*, ISBN 978-2-87487-007-1, pp. 554-557, Amsterdam, The Netherlands, Oct. 2008

Yang, B.; Skafidas, E. & Evans, R. J. (2008). Design of 60 GHz Millimetre-wave Bandpass Filter on Bulk CMOS, *IET Microwaves, Antennas & Propagation*, ISSN 1751-8725 (accepted for publication)

Yao, T.; Gordon, M.; Yau, K.; Yang, M. T. & Voinigescu, S. P. (2006). 60-GHz PA and LNA in 90-nm RF-CMOS, *Digest of Papers of the 2006 IEEE Radio Frequency Integrated Circuits Symposium*, 4 pages, ISBN 0-7803-9572-7, San Francisco, California, USA, June 2006, IEEE, Piscataway

Current Trends of CMOS Integrated Receiver Design

C. E. Capovilla and L. C. Kretly
School of Electrical and Computer Engineering, University of Campinas
Campinas, São Paulo, Brazil

1. Introduction

The use of CMOS technology for implementation of fully RFICs (Radio Frequency Integrated Circuits) is shown as a trend in the new devices for wireless communications. Nowadays, these circuits are found in many kind of applications, in which they can provide a lot of services including: Cellular Phones, Personal Mobile Service, Satellite, Specialized Radios (used by the Police, Fire-fighters, Emergency Services), and WLANs (Wireless Local Area Networks).

The potential of CMOS RFIC has been demonstrated in many academic works and by commercial devices. Due to the quick development of 3G and 4G technologies and the design of devices for systems that operate at standards such as WCDMA (Wideband Code Division Multiple Access), GSM/GPRS (Global System for Mobile Communication/General Packet Radio Service), WiMAX (Worldwide Interoperability for Microwave Access), and WiBro (Wireless Broadband), the CMOS RFIC has been more and more inserted, because of its good technical and commercial characteristics, becoming itself a challenge for the designers (Iniewski, 2007).

For modern and appropriated RFIC applications, new mobile systems demand antennas in small dimensions with wideband and reasonable gain, offering the possibility of a multiband operation (Liberti & Rappaport, 1999).

In addition, the UWB (Ultra WideBand) communication systems are an important advancement in wireless applications. They use a wide range of frequencies at very low power to transmit at high data rate. The low power allows these systems to use existing licensed RF bands without interfering with current users (Ismail & Gonzalez, 2006).

With the coming of the new digital standards, the data exchange over the wireless became predominant. In fact, in the 1990s, the GSM and IS-95 standards, evolved to include data transmission as an effective part of its services. Besides, the 3G and 4G are being applied and developed with voice and data integration. It has been foreseen that the data traffic will overtake the voice one. Furthermore, nowadays the costs of these devices and services for the data traffic are cheap enough to permit its continuous utilization into user's houses and offices.

In this way, the aim of this chapter is to provide a guide to the RF building blocks of smart communication receivers in accordance with the present state of the art. The goal is to show

the conception and development of several RFICs, for example, LNAs (Low Noise Amplifiers), mixer, and VCOs (Voltage Controlled Oscillators) in different applications.

The circuits presented here can supply the necessities for many mobile applications, in particular, for SMILE (Spatial MultIplexing of Local Elements) front-end receiver circuitry. As an example of a circuit developed for this technique, it is shown a multiplexed LNA with four channels to supply the necessity of multiplexing without losing the concern about noise or any other kind of design performance parameters (Capovilla et al., 2007). Page constraints have made it necessary to limit coverage in some areas to represents different areas as best as possible.

2. Smart receivers

The antenna array is one of the most promising techniques for increasing the system capacity in wireless communication. The demand for mobile systems emerges and the use of data transmission grows through applications of several protocols.

With the quick development of 3G and 4G technologies and the growth of the commercial applications for their equipments, the seeking for antenna technical solutions has increased a lot for these applications. Due to this fact, the antennas represent a fundamental role in its performance, strengthening this research area.

Using a variety of processing algorithms, usually managed by a DSP (Digital Signal Processor), the adaptive antennas adjust its radiation pattern dynamically to enhance the desired signal, null or reduce interference (Liberti & Rappaport, 1999). They are used to improve the received signals, minimizing interference and maximizing the desired receiving signal and to form dynamic beams.
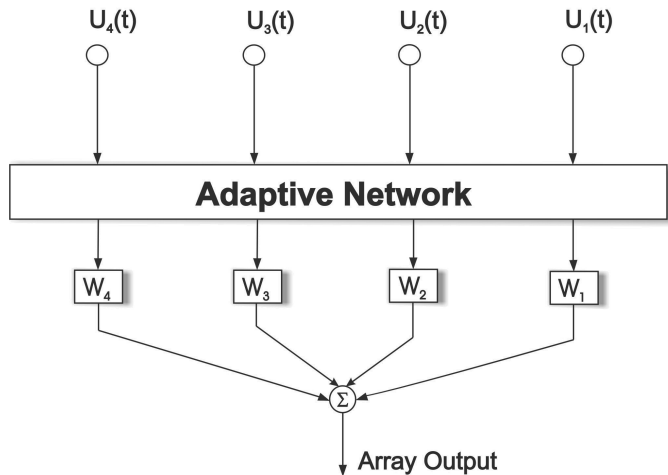

Fig. 1. – Adaptive array concept.

An adaptive system considers that the desired signal and the interfering one come from different directions. As can be seen in Fig. 1, to reduce the fading and the co-channel

interference, the system processes four input signals coming from different antennas of the array (u1(t), u2(t), u3(t), and u4(t)) to generate an output optimized signal.

This one is the work of the crossed correlation and of the relative signal levels between four received signals. The radiation pattern can be configured in real time through direct application of control algorithms as MUSIC (Multiple Signal Classification) (Ratnarajah & Manikas, 1998) and ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) (Roy & Kailath, 1989), which are examples of DOA (Direction Of Arrival) algorithms.

To estimate the best weight of the array (Godara, 1997a), efficient algorithms as LMS (Least Mean Squares) (Clarkson & White, 1987) and RLS (Recursive Least Squares) (Qiao, 1991) can be used. Due to this control over the radiation pattern envisaging a better management of the system, it is also possible to form dynamic cells using the multiple beams. This technique is known as SDMA (Space Division Multiple Access) (Godara, 1997b), and allows for different users the simultaneous operation of the same time/frequency slot, increasing the capacity of the system (Kuehner et al., 2001).
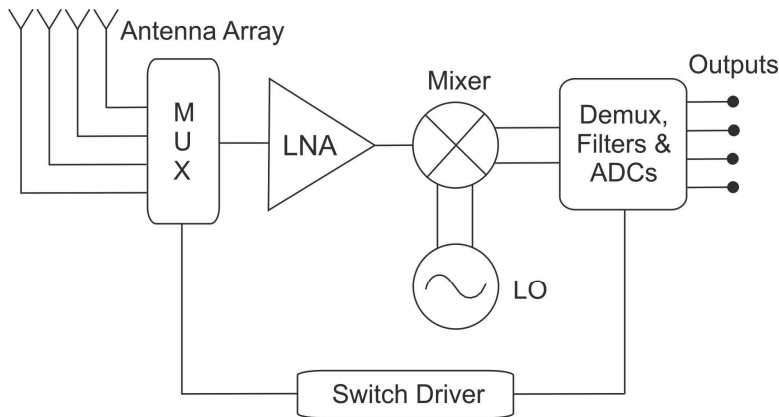

Fig. 2. – SMILE RF front-end receiver architecture.

Additionally, there are several techniques for optimization of smart antennas in spatial diversity. The DBF (Digital BeamForming) is one of those techniques that revolutionized the capabilities of antenna arrays. In the beginning, the DBF projects were motivated by military operations, however with the increasing interest in low cost WLAN, nowadays, there are studies in order to use the DBF in different applications. The DBF scheme provides a lot of advantages over analog beamforming including in hardware implementation (Doble & Litva, 1996).

For this one, the smart antenna array requires independent RF channels (RF switch, LNA, and mixer) for each array element, increasing hardware costs and the power consumption, which are proportional to the number of array elements. In this way, many efforts have been made aiming at reducing the use of repetitive RF channels. The works of Cheng (2001) and Ishii (2000) show some of the attempts in this direction, but only for limited functional environmental conditions.

The SMILE scheme is shown in Fig. 2. It appears as a new solution to solve these technical problems. This hardware technique reduces the RF channels of the smart receiver to only one without loss of signal fidelity. This functional characteristic is obtained by independently switching the array elements at a rate above the Nyquist frequency, according to the sampling theory. After processing the RF channel (through RF switch, LNA, and mixer), the spatially sampled signals are demultiplexed and low-pass filtered to form only one output (Fredrick et al., 2002).
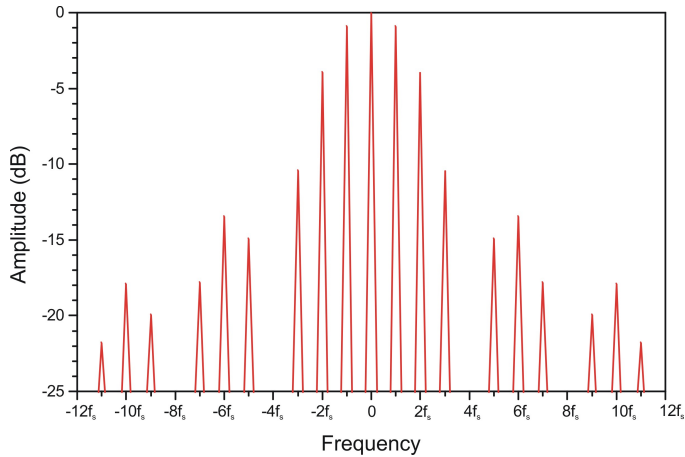
Fig. 3. – Baseband SMILE spectrum.

To test the scheme, a single-tone test was performed with an IF of 750kHz. After receiving the signal, the baseband spectrum of the multiplexed signal for the array rotate at 45° is shown in Fig. 3. Each of the four channels demultiplexed and recovered are shown in Fig. 4. The envelope shows the original data samples. This technique significantly reduces the RF hardware, getting the necessary functionality with only a fraction of the hardware requirements. Compared to N elements from a traditional system, the proposed system offers an N fold reduction in the RF hardware requirement also reducing the power consumption and the circuit size.

3. CMOS receivers

The RF basic blocks of receivers are composed by LNA, mixer, and LO (Local Oscillator). In this section, these circuits are presented and characterized in CMOS technology. Normally, at the foundry, the CMOS RF technology is derived from a process for manufacturing digital circuits, after a stabilization procedure, by adding masks and performing other slight modifications, such as, the use of thick metal technique for the top layer (Backer et al., 2001).
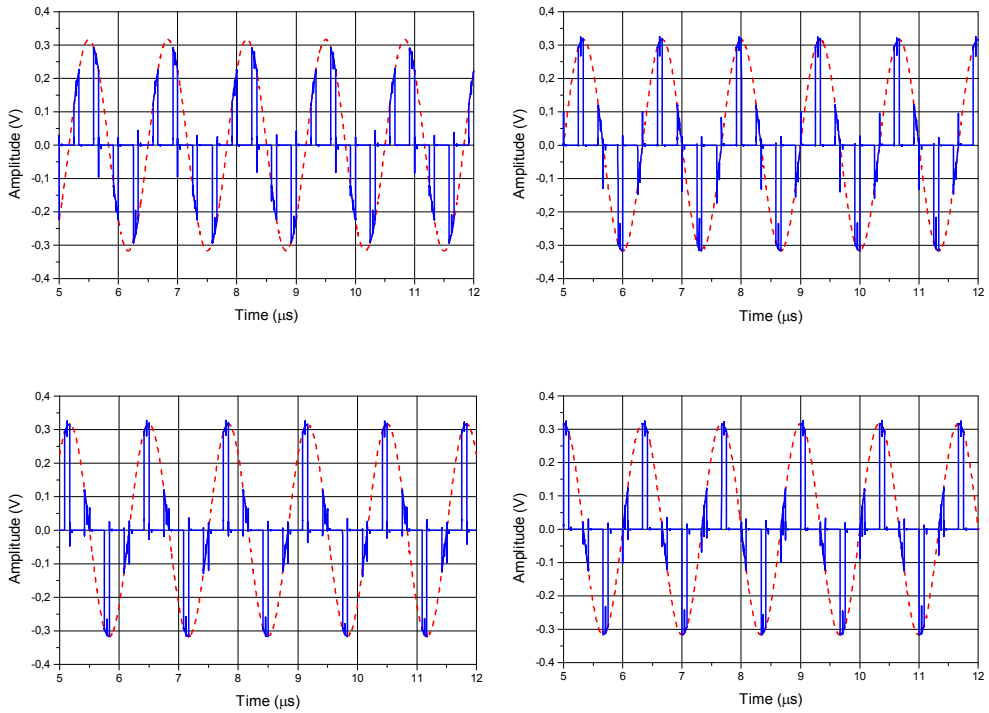
Fig. 4. – Recovered multichannel baseband data for array at 45°.

3.1 Low noise amplifier

The schematic of a basic common-source LNA is shown in Fig. 5. For simplicity, the bias network is represented only by Vbias and Rb (usually 5-10kΩ). The input and output are coupled with DC-block capacitors (not showed here). The use of inductive degeneration results in no additional noise generation since the real part of the input impedance does not correspond to a physical resistor. A mathematical representation of the noise from the whole amplifier circuit with neglected noise contribution of the transistor Mn2, is given by (Allstot et al.,2004):

$$F = 1 + \frac{\gamma}{\alpha}\frac{1}{Q}\left(\frac{\omega_0}{\omega_T}\right)\left[1 + \frac{\delta\alpha^2}{5\gamma}\left(1+Q^2\right) + 2\,|c|\sqrt{\frac{\delta\alpha^2}{5\gamma}}\,\right] \tag{1}$$

with:

$$Q = \frac{1}{\omega_0 C_{gs} R_s} \qquad\qquad \alpha \equiv \frac{g_m}{g_{d0}}$$

where: Q is the quality factor, g_{d0} is the drain conductance, and, γ, δ, c are fixed transistor parameters.
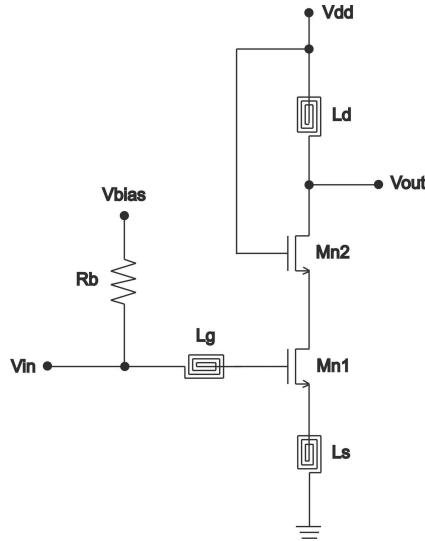

Fig. 5. – LNA schematic.

A simple analysis of the input impedance (Shaeffer & Lee, 2001) shows that:

$$Z_{in} = s\left(L_s + L_g\right) + \frac{1}{sC_{gs}} + \frac{g_m L_s}{C_{gs}} \qquad (2)$$

where: L_s and L_g are source and gate inductors, respectively, and g_m and C_{gs} denote small signal parameters of transistor Mn1 (C_{gd} and C_{ds} are neglected in this first-order approximation).

The input is matched to 50Ω by using inductors L_g and L_s, with the source inductor L_s chose to match the real part, and gate inductor L_g used to set the resonance frequency. Using some assumptions taken from the long-channel theory, the optimum width of the device Mn1 is given by (Lee, 1998):

$$W_{Mn1} \approx \frac{1}{3\omega_0 L_{eff} C_{ox} R_s} \qquad (3)$$

where: L_{eff} is the effective transistor length and C_{ox} is the oxide capacitance of the transistor. Equation 3 gives a definite width of the transistor, but for short channel devices, the CMOS technology leads to very large transistors. In this case, it is recommended to use multi-gates transistors to reduce the noise generated due to the resistance of the gate.

For the selection of the cascode transistor Mn2 width, two competing considerations should be made. The Miller capacitance of Mn1 can considerably reduce the gate and drain impedances of Mn1, degrading both the noise performance and the input matching.

This behaviour can be compensated by a large cascode device (C_g device), which reduces the gain of the C_s device. However, the parasitic source capacitance associated with a large C_g device increases the amplification of the C_g device. It was presented in publications that the ratio between C_s and C_g transistor widths varies from 0.5 (Guo & Hang, 2002), up to three (Goo & Dutton, 2002), based on simulations. Note that Mn2 also introduces noise in the amplifier, and that, the size of this transistor should be also constrained by the noise figure of the amplifier (Rafla & El-Gamal, 1999).
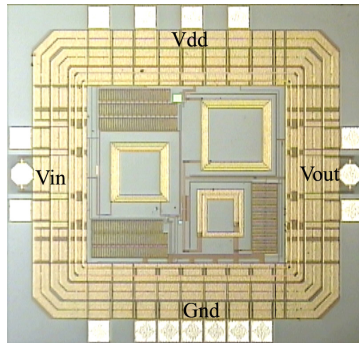

Fig. 6. – LNA die (1530 x 1425µm).

As an example, it is shown in Fig. 6 a LNA die with an area of 2.2mm^2. This circuit was fabricated by AMS (Austriamicrosystems) foundry with 0.35µm gate length and four metal layers (metal4 is a thick-metal layer used mainly in spiral inductors). The three spiral inductors are clearly visible. The input spiral inductors (L_g) and the input pads are at the left side of the die. The inductor at the lower right side is L_s and the one at the upper right side is L_d, which tunes the output of the LNA. The spiral inductors are fabricated with metal4 (thick-metal), which gives Q's of about eight. This value of Q is higher than a typical metal3 on-chip spiral inductor can provide (Li et al., 2004).

3.2 Mixer

The downconverter mixer translates an incoming RF signal to a lower frequency, being possible in this lower frequency to get necessary selectivity and gain for the receiver. A nonlinear device makes the multiplication of the RF signal and the LO signal in time domain. This multiplication results in output signals at sum and difference frequencies of the inputs. For selectivity reasons, the signal that always interests is the difference of the input signals and, usually, it is selected through a low-pass filter. Theoretically, to accomplish the frequency translation a nonlinear device with quadratic characteristic is used, but in implemented devices, this characteristic normally does not occur.
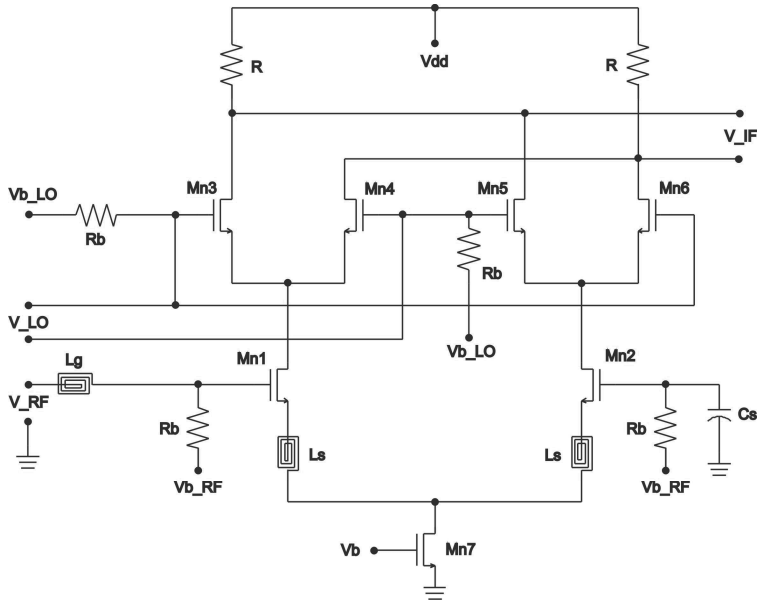
Fig. 7.– Double-balanced Gilbert cell as mixer.

So, if the nonlinear device presents N degree transfer characteristic, in the translation, other components will appear, and eventually, they can overlay the frequency of interest or to arise close to the same one, resulting in distortion. The CMOS mixers present some advantages if compared to bipolar ones. For instance, the transfer characteristic of the MOS transistor is approximately quadratic in saturation region, while the bipolar one is approximately exponential. Thus, the MOS transistor presents less harmonic distortion (Tsividis, 1999). Other advantage is that the MOSFET has better noise performance (internally generated noise). Looking through the topology, an advantage of the double-balanced structure in comparison with the single-balanced one is the good isolation between the LO and the IF port. Besides, there are other advantages such as the noise-rejection in common mode, better linearity, and less intermodulation (Lehne et al., 2000). In this way, for RFIC mixer, the Gilbert cell is the most common topology. Its choice usually is inevitable (Darabi & Chiu, 2005). Thus, due to the advantages and presented considerations, the researches with this type of mixer have been intensified in the last decade, resulting in modifications of classic structures and doing this topology almost unanimity in recent publications.

The mixer shown in Fig. 7 is a doubly balanced Gilbert cell with one arm of the RF differential pair connected to input and the other arm AC grounded. This differential input is widely used in CMOS downconverter mixers, since it provides a high impedance input to the low noise amplifier and is capable of driving a low impedance load at its output. To drive this one, the mixer output is buffered (not shown in the figure).

An example of implemented double-balanced mixer in 0.35μm CMOS technology is shown in Fig. 8. This prototype has an overall area of 2.45mm^2.
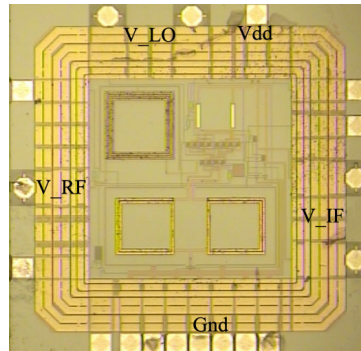
Fig. 8. – Double-balanced mixer die (1570 x 1560µm).

3.3 Local oscillator

The voltage controlled oscillator is a kind of oscillator in which the frequency of oscillation can be modified inside of a pre-determined band. Usually, there are three types of integrated VCO topologies: Ring oscillators, relaxation oscillators, and tuned oscillators (Razavi, 2001). The ring oscillators are implemented by digital inverter cells at feedback closed loop (odd number of inverters). Its integrated design is simple and compact. The frequency control is made through the current variation into the inverter cells, or eventually, for the modification of the inverter capacitance loads.

Its main intrinsic problem is the high phase noise due to the continuous switching of the inverters. So, its application for RFIC is not feasible (Backer, 2001), however it is indicated for many applications as, for example, the clock generation for digital or mixed-signal circuits.

Another topology is the relaxation oscillator, which works charging and discharging a capacitor with constant current. In the same way that the ring oscillator, its tuning is made by the modification of the current. Its easy integration and compact size become this topology attractive for integrated circuits, even though the high current consumption necessary to reduce the phase noise limits its RFIC applications.

The third usually integrated topology is the tuned oscillator, which contains a resonator LC tank or a tuned crystal. The resonator generates the oscillation and an active circuit supplies the energy necessary to compensate the resistive losses of the resonator. A difficulty that exists to integrate this type of oscillator is due to the low quality that still exists in the integrated passive devices, however the main problem of this type of oscillator is the large area of spiral inductors. But, there are lots of advantages in this topology, as steady-state oscillation, great spectral pureness, and low power dissipation. Because of these advantages, nowadays this type of oscillator is more frequently used in RFIC applications (Hajimiri & Lee, 2001). The tuned oscillator can be implemented in different topologies. For RFIC applications with differential design, normally is chosen the CMOS or NMOS LC due to its easy design. For single-end design, the Colpitts oscillator is more interesting than the Hartley oscillator due to the larger number of the spiral inductors used in the Hartley topology. The VCO shown in Fig. 9 is a NMOS LC with the frequency control performed through the variation of the capacitance (varactor) of the LC tank.
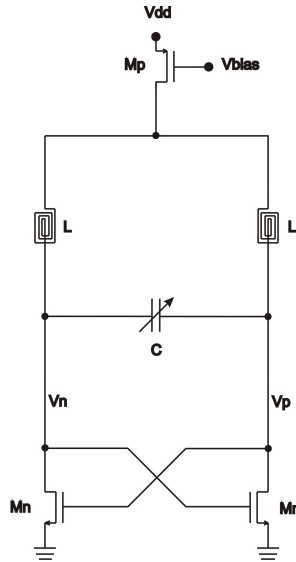
Fig. 9.– LC NMOS schematic.

This differential output is widely used in CMOS oscillators to make a direct connection with double-balanced mixers. If necessary to drive low impedance loads, the VCO output (Vp - Vn) must be buffered (not shown in the figure). An example of implemented oscillator in 0.35µm CMOS technology is shown in Fig. 8. This prototype has an overall area of 1.8mm^2.

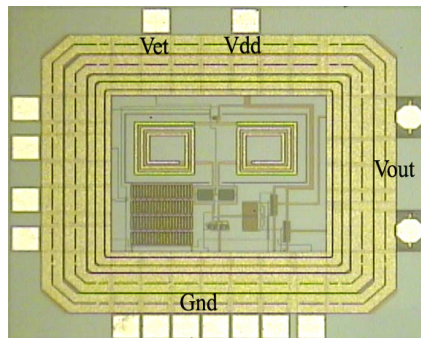Fig. 10. – LC NMOS die (1400 x 1280µm).

4. Implementation of an integrated SMILE receiver in CMOS technology

The use of antenna array in a smart system requires independent RF channels (LNA, mixer, etc) for each array element. This increases hardware costs and power consumption, which are proportional to the number of array elements. Another problem is that the multiple-feedline arrays and complex multiple RF circuits result in difficulties for optimized circuit integration. Also, more noise arises inside the system with the growing of electronic devices.

As the SMILE scheme is a front-end receiver architecture which uses only one RF channel, carrying multiplexed information from multiple antennas, in this section is shown, as an example of implementation for SMILE applications, a single RF channel using a fully integrated multiplexed LNA with four input channels, a double-balanced mixer, and a VCO as local oscillator. All circuits were fabricated in 0.35μm CMOS technology. The circuits are designed to operate at 2.5GHz band with an IF of 750kHz. If compared to a single channel system, it has the same performance with the addition of switching functionality. The proposed system presents power consumption four times less and an overall area reduction around 70% when compared to a conventional smart antenna architecture using four separate RF channels.

4.1 Multiplexed low noise amplifier

The smart antenna systems, generally, are composed of separate LNAs, and, in this case, the LNAs are always polarized, what generates high power consumption with poor power efficiency, due to only one channel to be used in each time slot. This problem grows, becoming critical, with the increase of the channels number. As an interesting hardware solution, the SMILE technique is implemented with the use of a novel type of LNA, which compared to a single LNA, has the same performance with the addition of switching functionality (Capovilla et al., 2007).
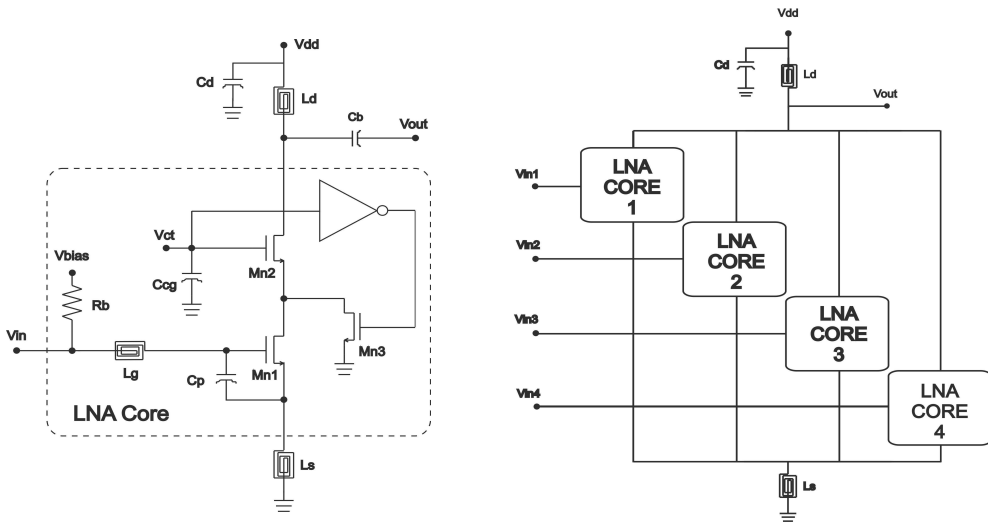

Fig. 11. – Multiplexed LNA with four input channels.

Analyzing this circuit shown in Fig. 11, the proposed structure is composed by four LNA-Core in parallel sharing the same L_d and L_s spiral inductors. Internally, the LNA has a capacitance of 1pF (C_d) from V_{dd} to ground for RF decoupling of the DC supply line.

The common-gate stage works like an NMOS switch and the shunt transistor Mn3 is used to improve the isolation of the LNA-Core. This shunt transistor is activated by a signal generated in the inverter cell that is activated with the same control signal that activates the

Mn2. When it is turned "ON", any residual RF signal existing in the LNA cascode pair is grounding (Huang & O, 2001).

A prototype of the LNA was designed to operate at 2.5GHz band (American WCDMA and WiMAX applications). A select pin controls which input is active by steering the current through the selected input stage and cutting off from the others.

This provides the optional switching functionality in front of the LNA, but without the insertion loss and noise figure penalty from the switch. In each channel, a control circuit is connected directly to the gate of the transistor Mn2.

When the control signal indicates an operation of a specific channel, the Mn2 of this channel is switched "ON" and the Mn3 "OFF". Conversely, for the other channels, the Mn1 is switched "OFF" and the Mn3 "ON". Therefore, only a desired channel is conducted to the output.
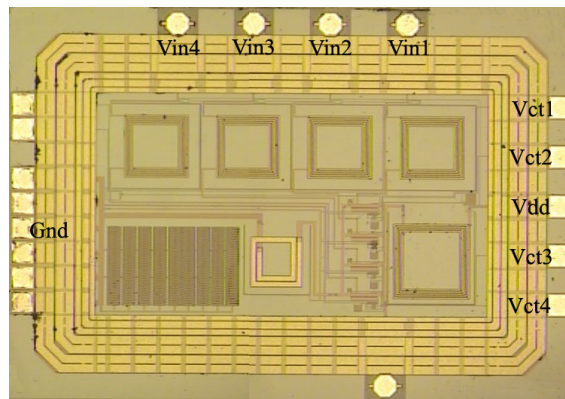

Fig. 12. – Multiplexed LNA with four channels die (2230 x 1580μm).

A photograph of the multiplexed LNA die is shown in Fig. 12 and was fabricated with an overall area of 3.52mm2. The six spiral inductors are clearly visible. The input spiral inductors (L_g) and the input pads are at the upper side of the die. The spiral at the lower left is Ls and the one at the lower right is L_d, which tunes the output of the LNA.

For comparison purposes a separate LNA, like a LNA-Core, has been fabricated having an area of 2.2mm2. The proposed LNA presents an area reduction of about 60% and power consumption four times less than a conventional system using four separate amplifiers.

4.2 Integrated single RF channel receiver

The RF channel of this smart receiver is composed by the fully integrated multiplexed LNA and a double-balanced mixer. As local oscillator is used an integrated LC NMOS VCO. The implemented system design in a SMILE scheme is shown in Fig. 13.
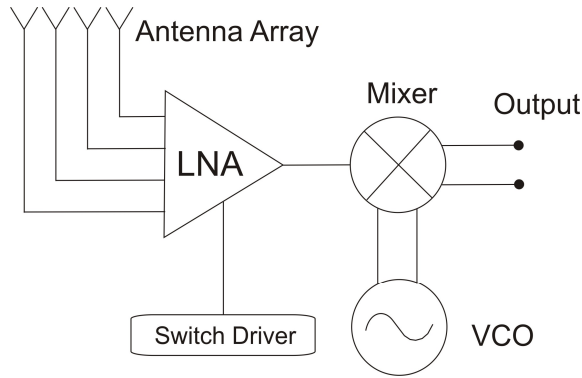
Fig. 13. – SMILE receiver scheme.

The proposed solution also reduces the overall noise figure of the system, since no RF switches are used. To test the circuits, it is used COB (Chip-on-Board) packaging technique, in which the bare die is directly bonded to the PCB (Printed Circuit Board). The board is fabricated using FR-4 material (ε_r=4.4 and thickness 1.6mm). It was prepared for characterization by using a microstrip test fixture.

All measurements are made with a 3.3V supply voltage and the current consumptions, under nominal conditions, are 6.5, 12, and 8mA for the LNA, mixer, and VCO, respectively. The total consumption is four times less than a normal smart antenna receiver with the same characteristics.
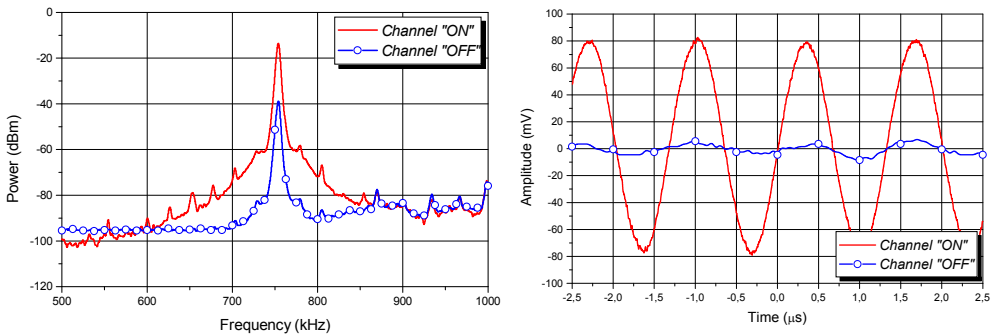

Fig. 14. - Switching characteristics of the channel one.

For these measurements, the RF power is -22dBm (2.6GHz) with a LO power of -2dBm (2.599250GHz). In Fig. 14 is shown the IF spectrum and IF waveform output in two situations: when the control signal 1 is switched "ON" and the other controls are switched "OFF" (channel one "ON"); and when the control signal 2 is switched "ON" and the other controls are switched "OFF" (channel one "OFF"). The total gain from channel 1 to IF output is above 9dB with an "ON-OFF" ratio of more than 25dB, sufficient for modern wireless applications.

The measured parameters of the circuits are summarized in Table 1. The measured SSB NF (Single-Side Band Noise Figure) of the mixer is 9.3dB and the LNA NF is 2.4dB, which results in a total NF of 3.9dB. The presented NF values are due to the measured gains to be lower than the design ones. This lower gain is due to the process variation to present a thicker gate-oxide in these fabricated prototypes. The total third-order nonlinearity is mainly limited by the transconductance stage of the mixer, with a total value of -2.8dB.

	S_{11}	Gain	Isolation	IIP3
	(dB)	(dB)	(dB)	(dBm)
LNA	-10	5	-17	4.5
Mixer	-34	4	-	3
Total	-12	9	-17	-2.8

Table 1. - Measured parameters of the circuits.

The SMILE scheme can also be used in quadrature system and, for this kind of application, a quadrature oscillator is needed. A QVCO generates two components in quadrature that feed the switching transistors of two mixers. One for the signal in phase and the other for the signal in quadrature. The design of the receiver is approximately equal to the same impedance and bias matching of the single design. The only difference is that the LNA output normally drives one mixer, and in this case, it drives two mixers in parallel. A simple impedance matching is enough to the design in quadrature works perfectly.

5. Conclusion

The application of CMOS RFIC has been shown as a trend for the implementation of modern wireless communications. The use of smart antennas to these systems together with the optimization of the receiver can result in many novel mobile applications. In scope of this chapter, the reconfigurable terminals using the smart antennas are considered as interesting environments to apply the RFIC CMOS. However, with the increase of the number of channels to improve the efficiency of the system, the RF devices have increased at same rate. For an array with N antennas, the total number of RF channels required is N. In this way, the power consumption and the devices expense is approximately N times those in a single antenna.

Another problem is about the hardware interconnectivity that becomes more and more complex with the increase of the channels. The SMILE technique appears as a new solution that must be able to solve this problem. The objective of this technique is to reduce the number of needed RF channels to only one, without loss of signal fidelity. This is obtained by independently switching the array elements at a rate above the Nyquist frequency, according to the sampling theory. After processing the RF channel, the spatially sampled signals are multiplexed to form only one output. Therefore, the aim of this chapter was to provide a simple guide to the building blocks of communication receivers in accordance with the present state of the art for this smart receiver application. It was shown the conception of several RFICs, for example, LNAs, mixer, and VCO implemented for the use in the stage of frequency conversion. All the circuits shown in this chapter, besides their

research characteristics, can be easily and quickly implement on commercial systems. As a final circuit example, an integrated single RF channel of the SMILE receiver prototype was shown. This receiver uses four multiplexed channels coming from four antennas. The multiplexation is made by a multiplexed LNA exclusively developed for this project. The characterization of the multiplexed amplifier, according to the "ON/OFF" functionality, confirms the good performance of the design operating at 2.5GHz band.

An early attempt at designing RFIC in the semiconductor foundry technologies using 0.13μm or nanometer process has revealed a new paradigm for the designers. A successful design approach for highly integrated RFIC in this environment will grow a lot, and in Brazil, the contact with TSMC (Taiwan Semiconductor Manufacturing Company) for migration of the circuits showed here to nanometer technologies are occurring at this moment. With all these attractive and perspective characteristics, the SMILE technique and the CMOS RFIC circuits, which have been separately or jointly developed for this application, must find great applications in modern wireless communications systems.

6. Acknowledgments

7. References

Allstot, D. J., Li, X. & Shekhar, S. (2004). *Design considerations for CMOS low noise amplifiers*, IEEE Radio Frequency Integrated Circuits Symposium.

Baker, J., Li, H. W. & Boyce, D. E. (2001). *CMOS Circuit Design, Layout, and Simulation*, IEEE Press.

Capovilla, C. E., Barbin, S. E. & Kretly, L. C. (2007). *A 2.6 GHz Integrated CMOS Low Noise Amplifier with Multiple Inputs for SMILE Array Receiver Application*, APMC 07 - IEEE - Asia-Pacific Microwave Conference - Bangkok - Thailand.

Clarkson, P. M. & White, P. R. (1987). *Simplified analysis of the LMS adaptive filter using a transfer function approximation*, IEEE Transactions Acoustic, Speech and Signal Processing 35: 987–993.

Darabi, H. & Chiu, J. (2005). A *noise cancellation technique in active RF-CMOS mixers*, IEEE Journal of Solid-State Circuits 40(12): 2628–2631.

Doble, J. & Litva, J. (1996) *Digital Beamforming in Wireless Communications*, Artech House.

Fredrick, J. D., Wang, Y. & Itoh, T. (2002). *A smart antenna receiver array using a single RF channel and digital beamforming*, IEEE Transactions on Microwave Theory and Techniques 50(12): 3052–3058.

Godara, L. C. (1997a). *Applications of antenna array to mobile communications*, part II: Beamforming and direction-of-arrival considerations, Proceeding of the IEEE, Vol. 85, pp. 1195– 1245.

Godara, L. C. (1997b). Applications *of antenna array to mobile communications*, part I: Performance improvement, feasibility and system considerations, Proceeding of the IEEE, Vol. 85, pp. 1031–1060.

Goo, J. & Dutton, R. W. (2002). *A noise optimization technique for integrated low-noise amplifiers*, IEEE Journal of Solid-State Circuits, vol. 37, no. 08, pp. 994–1001, August.

Guo, W. & Hang, D. (2002), *Noise and linearity analysis for a 1.9 GHz CMOS LNA*, ICMMT - International Conference on Microwave and Millimeter Wave Technology, pp. 923–927, August.

Hajimiri, A. & Lee, H. (2001). *The design of low noise oscillators*, Kluwer Academic Publishers.

Huang, F. J. & O, K. K. (2001), A 0.5-µm CMOS T/R switch for 900-MHz wireless applications, IEEE Journal of Solid-State Circuits, vol. 36, no. 03, pp. 486–492, March.

Iniewski, *K. (2007). Wireless technologies: circuits, systems, and devices*, CRC Press

Ismail, M. & Gonzalez, D. R. (2006). *Radio Design in Nanometer Technologies*, Springer.

Kuehner, R., Todd, T. D., Shad, F. & Kezys, V. (2001). *Forward-link capacity in smart antenna base stations with dynamic slot allocation*, IEEE Transactions on Vehicular Technology 50(04): 1024–1038.

Lee, T. H. (1998). *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press.

Lehne, M., Stonick, J. & Moon, U. (2000). *An adaptative offset cancellation mixer for direct conversion receivers in 2.4GHz CMOS*, IEEE I. S. of Circuits and Systems 01: 319–322.

Li, Z., Quintal, R. & O, K. K. (2004). *A dual-band CMOS front-end with two gain modes for wireless LAN applications*, IEEE Journal of Solid-State Circuits, vol. 39, no. 11, pp. 2069–2073, November.

Liberti, J. C. & Rappaport, T. S. (1999). *Smart Antennas for Wireless Communications IS-95 and Third Generation CDMA Applications*, Prentice-Hall International.

Petrus, P. (1997). *Novel Adaptive Array Algorithms and Their Impact on Cellular System Capacity*, Ph.D. thesis, Faculty of the Virginia Polytechnic Institute.

Qiao, S. (1991). *Fast adaptive RLS algorithms: A generalized inverse approach and analysis*, IEEE Transactions Signal Processing 39: 1455–1459.

Rafla, R. & El-Gamal, M. (1999). *Design of a 1.5V CMOS integrated 3GHz LNA*, IEEE International Symposium on Circuits and Systems, vol. 02, pp. 400–443, June 1999.

Ratnarajah, T. & Manikas, A. (1998). *An H1 approach to mitigate the effects of array uncertainties on the music algorithm*, IEEE Signal processing letters 05(07): 185–188.

Razavi, B. (1998). *RF Microelectronics*, Prentice-Hall

Razavi, B. (2001). *Design of analog CMOS integrated circuits*. McGraw-Hill International.

Roy, R. & Kailath, T. (1989). *Esprit estimation of signal parameters via rotational invariance techniques*, IEEE Transactions Acoustic, Speech and Signal Processing 37: 984–995.

Shaeffer, D. K. & Lee, T. H. (2001). *The Design and Implementation of Low-Power CMOS Radio Receivers*, Kluwer Academic Publishers.

Tsividis, Y. (1999). *Operation and Modeling of the MOS Transistors*, 2th ed., McGraw-Hill.

Power Amplifier Design for High Spectrum-Efficiency Wireless Communications

Steve Hung-Lung Tu, Ph.D.
Fu Jen Catholic University
Taiwan

1. Introduction

The growing market of wireless communications has generated increasing interest in technologies that will enable higher data rates and capacity than initially deployed systems. The IEEE 802.11a standard for wireless LAN (WLAN), which is based on orthogonal frequency division multiplexing (OFDM) modulation, provides nearly five times the data rate and as much as ten times the overall system capacity as currently available 802.11b wireless LAN systems (Eberle et al., 2001; Zargari et al., 2002; Thomson et al., 2002). The modulation format of the IEEE 802.11a is OFDM (Orthogonal Frequency Division Multiplexing) which is not a constant-envelope modulation scheme; more sensitive to frequency offset and phase noise, and has a relatively large peak-to-average power ratio. These reasons induce the linearity requirements, which are crucial for power amplifier design. Among various linearization techniques, transistor-level predistortion is the simplest approach to implement and can be realized in a small area, which makes it be the most compatible with RFIC implementation.

Conventionally, WLAN has been implemented with multi-chip approach or single chip with processes other than CMOS. For example, the radio frequency (RF) and intermediate frequency (IF) sections are fabricated with GaAs or BiCMOS processes, and the baseband DSP section with a CMOS process. Note that the complexity and cost will be dramatic in a wireless LAN system with hybrid processes, which makes the CMOS process be the most promising approach to achieve high integration level, low power consumption, and low cost for the integration of baseband and RF front-end circuits of a WLAN system.

There are two different modulation schemes employed in wireless communication standards: linear modulation and nonlinear modulation, in which the former one is employed in North American Digital Cellular (NADC) standard whereas the latter one is also called constant-envelope modulation which employed in the European Standard for Mobile Communications (GSM). The main requirements for power amplifiers (PA's) employed in wireless communications are generally high power efficiency and low supply voltage operating at high frequencies. Class-E PA's have demonstrated the potential of high power efficiency whereas due to the operation characteristics, it can only be adopted in constant-envelope modulation applications. The linear modulation scheme, on the other hand can achieve high spectrum efficiency, which is especially suitable for the application of

wireless communications. A power amplifier that can achieve high power efficiency while providing high spectrum efficiency is therefore highly desired. To discuss this issue, in this chapter a class-AB type amplifier in a standard CMOS process is investigated together with the presentation of a transistor-level predistortion compensation techniques.

The main theme of this chapter is aimed at providing the fundamental background knowledge concerned with linear PA design for high spectrum-efficiency wireless communications. Nevertheless, we also present the design considerations of the state-of-the-art linear PA's together with the design techniques operating at the gigahertz bands in CMOS technologies. To conclude the chapter, we investigate a design and implementation of a Class-AB PA operating at GHz for IEEE 802.11 wireless LAN to demonstrate the feasibility.

2. Design Concept of CMOS Power Amplifiers

Conjugate matching is fully understood as making the value of load resistance equals to the real part of the generator's impedance. Since the maximum power will be delivered to the load, however, this delivering power will be limited by the maximum rating of the transistor. This phenomenon can be shown in Fig. 1. For utilizing the maximum current and voltage swing of the transistor, a lower than the real part of generator's impedance is chosen for maximum power transformation.
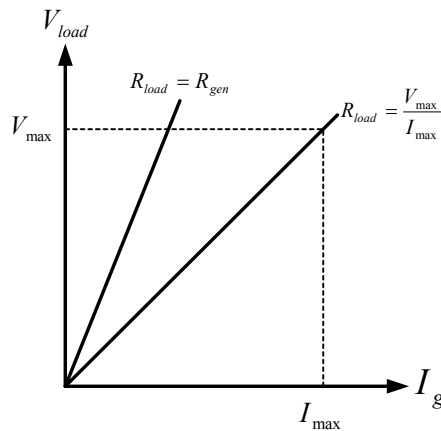

Fig. 1. Conjugate Matching and Power Matching

The load-line match represents an actually compromise that extracts the maximum power from the power devices, and simultaneously maintains the output swing within the limitation of the power devices and the available DC supply. In a typical situation, the conjugate matching yields a 1-dB compression power about 2dB lower than that can be attained by the correct load line matching (power matching), which means the power device can deliver 2dB lower power than the manufactures specify. So the power matching condition has to be taken seriously, despite the fact that the gain of the PA circuits is lower than conjugate matching at lower signal levels.

Another design concept of CMOS power amplifiers is the knee voltage effect of deep sub-micron CMOS transistors. The knee voltage (pinch-off voltage) divides the saturation and linear operation region of the transistor. Typically, for a power transistor may be 10% or 15% of the supply voltage, and the optimum load impedance is

$$R_{opt} = \frac{V_{max} - V_{Knee}}{I_{max}} \tag{1}$$

Notice that the knee voltage can be as high as 50% of the supply voltage for deep sub-micron CMOS technologies as shown in Fig. 2. Therefore, preventing the CMOS transistor from operating in the linear region doesn't result in the optimum output power. Also, both the saturation and linear operating regions must be considered in determining the optimum output load impedance since a lot portion of RF cycle could be in the linear operating region.
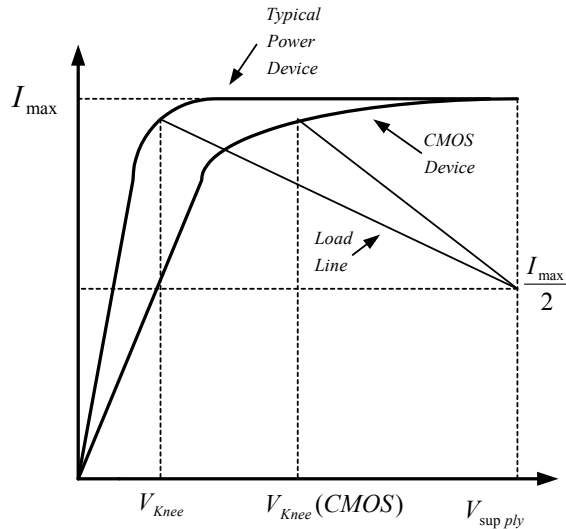

Fig. 2. Knee Voltage of Typical Power Device and CMOS device.

Another issue is the choice of device size of each amplifying stage. A simple class-A amplifier can briefly explain this issue as shown in Fig.3, in which RFC means radio frequency choke with large impedance compared with load impedance R_L.
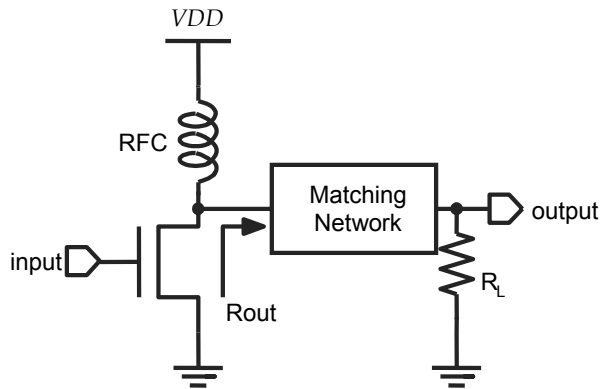
Fig. 3. Simple circuit of class-A amplifier.

Load impedance R_L is generally equal to 50 ohms and the matching network is tuned to obtain the device maximum output power. When the device output power is reaching to the maximum, the output impedance $Rout$ is defined as the optimum load impedance $Ropt$. In a class-A amplifier, the device plays a role of a voltage-dependant current source as shown in Fig.4, in which I_{max} is the maximum available current of the device, I_{min} is the minimum current of the device. V_{max} is the maximum tolerance voltage of the device between drain and source of the device. V_{min} is the knee voltage of the device. V_{dc} and I_{dc} are the DC bias of the device. Therefore, the device voltage swing and current swing are V_{max}-V_{min} and I_{max}–I_{min}, respectively.
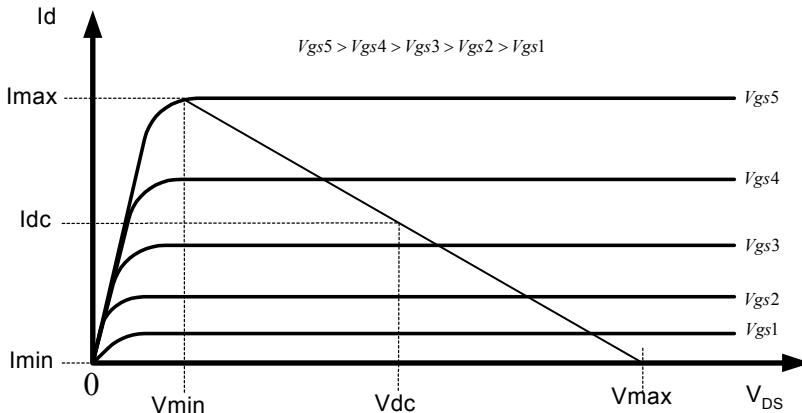

Fig. 4. I-V Curve of a NMOS device.

The optimum load impedance $Ropt$ for maximum AC swing can thus be described as

$$R_{opt} = \frac{V_{max} - V_{min}}{I_{max} - I_{min}} \tag{2}$$

and the output power of the device can be expressed as

$$P_{RF} = V_{rms} I_{rms} = \frac{V_{max} - V_{min}}{2\sqrt{2}} \cdot \frac{I_{max} - I_{min}}{2\sqrt{2}} = \frac{(V_{max} - V_{min})(I_{max} - I_{min})}{8} \tag{3}$$

Since DC power consumption is

$$P_{DC} = \frac{V_{max} + V_{min}}{2} \cdot \frac{I_{max} + I_{min}}{2} = \frac{(V_{max} + V_{min})(I_{max} + I_{min})}{4} \tag{4}$$

the power efficiency η is thus given by

$$\eta = \frac{P_{RF}}{P_{DC}} = 0.5 \frac{(V_{max} - V_{min})(I_{max} - I_{min})}{(V_{max} + V_{min})(I_{max} + I_{min})} \tag{5}$$

Theoretically, V_{min} and I_{min} are zeros, and ideal drain efficiency of a class-A amplifier is 50%. However, in fact V_{min} and I_{min} are not equal to zeros, which implies that the drain efficiency should be less than 50%.

3. Power Amplifier Linearization Techniques

Feedback linearization techniques are the most general approaches employed in RF power amplifier design such as in the North American Digital Cellular (NADC) standard, a CMOS power feedback linearization is employed to linearize an efficient power amplifier transmitting a DQPSK modulated signal (Shi & Sundstrom, 1999), in which a reduction of more than 10dB in the adjacent channel interference was achieved according to the experimental results.

Fig. 5 shows a PMOS cancellation transistor-level linearization technique (Wang et al., 2001). The measurement results demonstrate that the amplifier with nonlinear input capacitance compensation has at least 6-dB IM3 (Third-order intermodulation intercept point) improvement in a wide range of output powers compared with the non-compensated amplifier whereas the disadvantages are low power gain and increasing input capacitance.
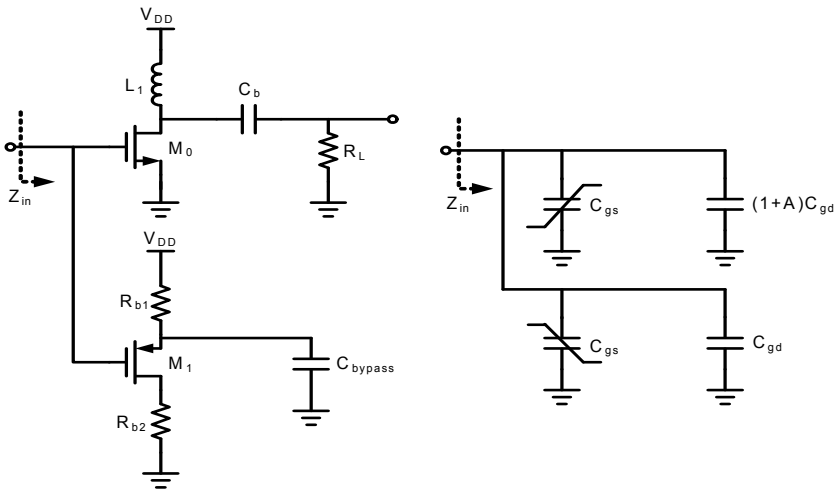
Fig. 5. Transistor-level linearization techniques – PMOS cancellation.

A miniaturized linearizer using a parallel diode with a bias feed resistance in an S-band power amplifier was also proposed (Yamauchi st al., 1997). The diode linearizer can improve adjacent channel leakage power of 5dB and power-added efficiency of 8.5%. Note that the improvement is based on 32 kb/ps, $\pi/4$ shift QPSK modulated signal at 28.6 KHz offset with a bandwidth of 16 KHz.

A miniaturized "active" predistorter using cascode FET structures was also applied to linearize a 2-GHz CDMA handset power amplifier. The ACPR (Adjacent Channel Power Ratio) improvement of 5dB was achieved (Jeon et al., 2002). Unlike the previously reported predistorters, this "active" predistorter can provide 7 to 17-dB gain which alleviates the requirement of additional buffer amplifiers to compensate the loss of the predistorter.

Another transistor-level linearization technique using varactor cancellation is shown in Fig.6 (Yu et al., 2000), which the approach improves 10-dB spectral regrowth with a low loss at 2GHz. However, the GaAs FET amplifier has AM-PM distortion under large-signal operating conditions due to the non-linear gate-to-source capacitance C_{gs} and the disadvantages are high cost, low integration with other transmitter circuits, and occupy a large PCB footprint.

A complex-valued predistortioner chip in CMOS for baseband or IF linearization of RF power amplifiers has been implemented (Westesson & Sundstrom, 1999). By choosing the coefficients for the predistortion polynomial properly, the lower-order distortion components can be cancelled out. Results of measurement performed as two-tone tests at an IF of 200MHz with 1MHz tone separation, using the chip for linearization gives a reduction of IM3 and IM5 with more than 30 and 10dB, respectively
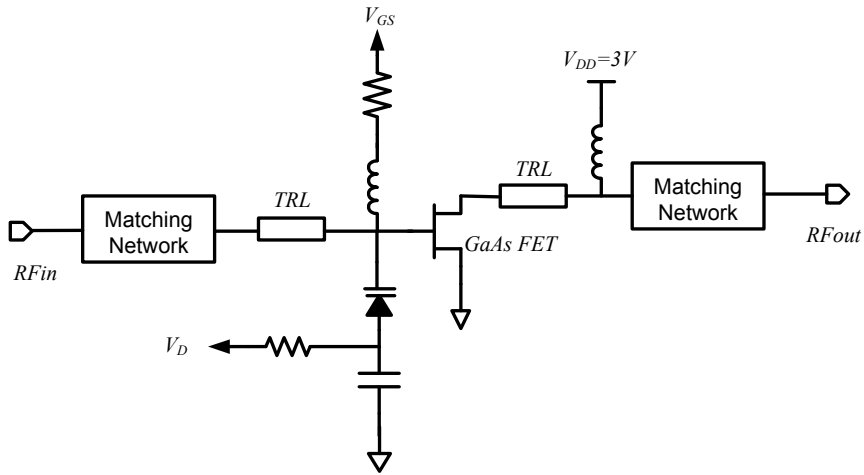
Fig. 6. Transistor-level linearization techniques – varactor cancellation.

Digital predistortion is a technique that counteracts both adjacent channel interference and BER degradation of power amplifiers. By employing digital feedback and a complex gain predistortion present, the experimental results demonstrate that a reduction in out of band spectra in excess of 20dB can be achieved (Wright & Durtler, 1992).

4. Predistortion Techniques for Linearization

Predistortion techniques are popular approaches for linearity improvement in power amplifier design. The concept is placing a black box on the PA input, which consumes little power and provides an acceptable linearity improvement instead of employing more complex circuitry to enhance system linearity. Basically, all predistortion approaches are open loop and can only achieve the level of linearization of closed-loop systems for limited periods of time and dynamic range. Recent research focuses on predistortion techniques offered by DSP. The basic concept is shown in Fig.7, where a predistorter preceding the nonlinear RF power amplifier implements a complementary nonlinearity, such that the combination of the two nonlinearities results in a linearized output signal. In practice, the lower orders nonlinear terms, such as third and fifth, is the most troublesome in communication applications. Even in practical PA models that consist of a couple of lower order nonlinear polynomial terms cannot be accurately estimated.
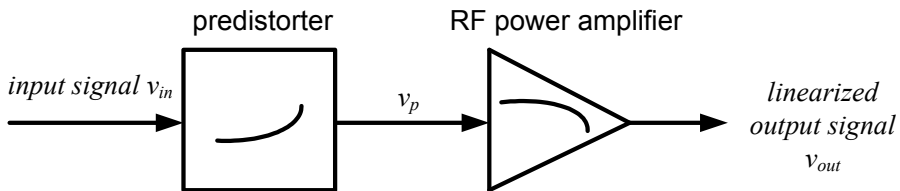

Fig. 7. Concept diagram of predistortion linearization.

4.1 Analog predistorters

Analog predistorters can be classified into two categories: 'simple' predistorters and 'compound' predistorters. The simple predistorters comprise one or more diodes, and the compound predistorters synthesize the required nonlinear characteristic using several sections to compensate different degree of distortion.

Simple analog predistorters mainly use a nonlinear resistive element such as a diode or an FET device as an RF voltage-control resistor that can be configured to provide higher attenuation at low drive levels and lower attenuation at high drive levels. A simple predistorter linearized RF power amplifier has been developed for 1.95-GHz wide-band CDMA (Hau et al., 1999), in which the amplifier is based on a heterojunction FET and its linearity and efficiency are improved by the employment of a MMIC simple analog predistorter which is shown in Fig.8. Gain expansion is observed when V_c is lower than –1V. Insertion loss (IL) is less than 5dB for a gain expansion of 2dB. Phase compensation was obtained from the MMIC predistorter as a result of the use of two inductors.
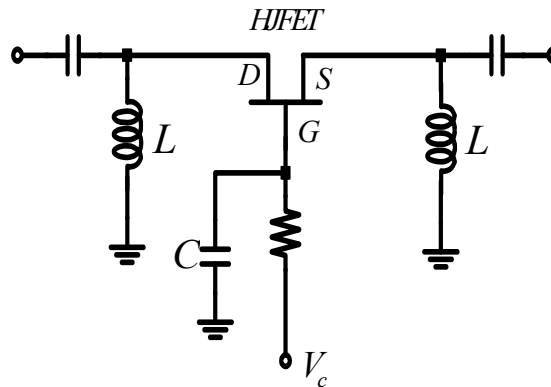

Fig. 8. Schematic of the MMIC predistorter.

The block diagram of a compound cuber predistortion system is shown in Fig.9, in which the input signal is split into two paths, and recombined in 180° phase shift at the output preceding PA (Morris & McGeehan, 2000). The key point of cuber predistorter is that the distortion terms can be scaled and phase shifted independently from the original undistorted input signal. Since the out of phase path can be set only for the third-order term, only the distortion term can be cancelled. For the reasons, this system is sometimes called a "cuber". However, there is a significant insertion loss in the combiner and splitter. Note that the lower coupling factors into and out of the cuber will result in a few losses in the main path. The independent two paths for high levels of IMD correction need a good gain and phase match.
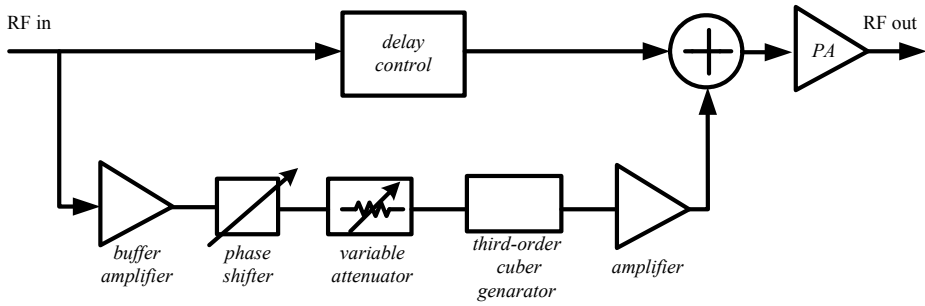
Fig. 9. Block diagram of a compound cuber predistorter.

4.2 DSP predistortion techniques

This approach is attractive since most modern radio frequency transceivers employ some form of DSP in their baseband processing as illustrated in Fig. 10.
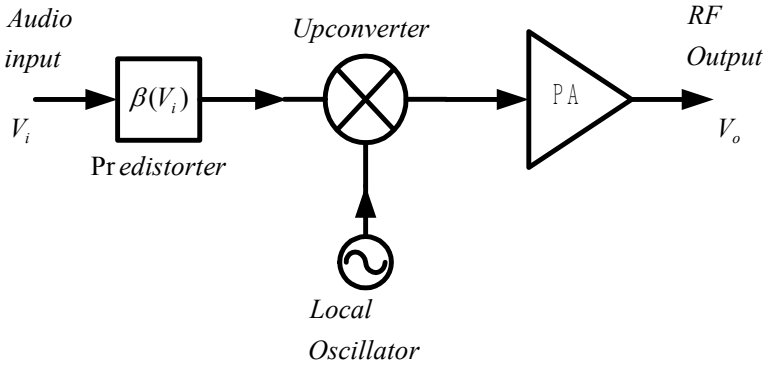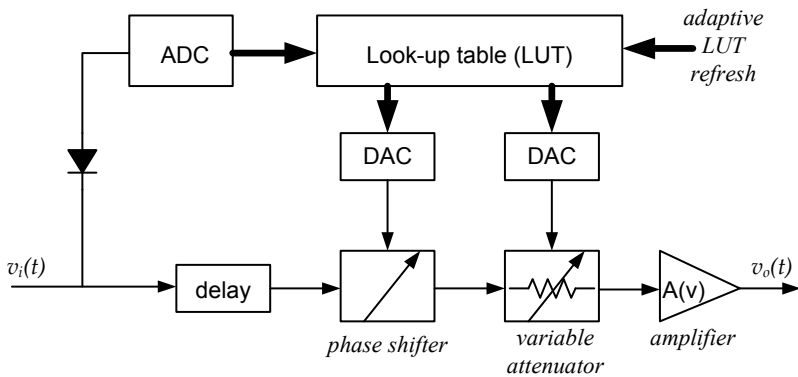

Fig. 10. Baseband predistortion system.

Fig. 11. DSP look-up table predistortion scheme.

A DSP look-up table predistortion system illustrates in Fig. 11. It should be noted that the system employs an input signal delay element to compensate the processing delays in the detection and DSP signal processing. The main limitation of the scheme is the speed of the detection and DSP itself.

The correction signals contain multiple harmonics of the baseband signal in order to perform the necessary predistortion function, which imposes a stringent requirement on the data converters. The precision of the look-up table is an important issue, which it can be implemented either physically or by a suitable algorithm. Moreover, the envelope input sensing is also a difficult task when the input signal throughputs continue rising. Note that a trade-off between the precision of detection process and the number of RF cycles employed to determine the final detector output is existed for the classical envelop detectors.

5. Linearity Improvement Circuit Techniques

Modern communication standards employ bandwidth-efficient modulation schemes such as non-constant envelope modulation techniques to prevent spectral re-growth problem, AM-AM, and AM-PM distortions, which means that some extra circuits for linearization purpose in power amplifier design are required.

Nevertheless, employing a linear PA's is a straightforward approach whereas it is also an inefficient method to meet the requirement of linearity. By taking advantage of the characteristics of high efficiency and applying some linearization techniques, nonlinear PA's may be a promising alternative. In this section, we investigate two transistor-level linear techniques to improve linearity of CMOS PA's namely, one is the nonlinear capacitance compensation scheme and the other is a parallel inductor compensation scheme. These two approaches will be described in the following subsections.

5.1 Nonlinear capacitance compensation technique

A deep sub-micron MOSFET RF large signal model that incorporates a new breakdown current model and drain-to-substrate nonlinear coupling is shown in Fig. 12 (Heo et al., 2000). This model includes a new breakdown current I_{dsB} with breakdown voltage turnover behavior and a new nonlinear coupling network of a series connection of C_{dd} and R_{dd} between the drain and a lossy substrate. The robustness of the new nonlinear deep sub-micron MOSFET model has been verified through load-pull measurements including IMD and harmonics at different termination impedance and bias conditions.
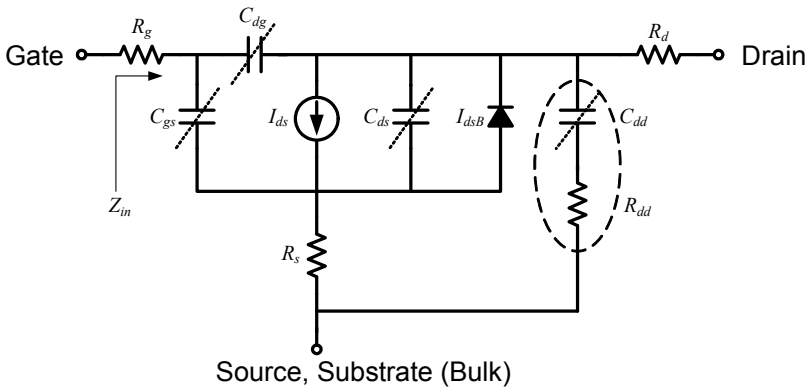
Fig. 12. Equivalent nonlinear model of a deep sub-micron NMOS device (slashed components are bias dependent).

A nonlinear capacitance cancellation technique to cancel the bias dependent input capacitance of the amplifier has been proposed and a prototype single-stage amplifier with a measured drain efficiency of 40% and a power gain of 7dB at 1.9GHz was reported in (Wang et al., 2001). The measured results indicate that the amplifier with nonlinear capacitance compensation has at least 6-dB IM$_3$ improvement in a wide range of output powers compared with the original amplifier without compensation.

The idea of the nonlinear capacitor compensation technique is that during the drain current clipping when the input signal is large enough to turn device 'on' and 'off', which the dramatical change in C_{GS} will generate distortion since Z_{in} is not keeping constant in signal amplification. The input impedance of the amplifier is approximately (ignore the R_S)

$$Z_{in} = \frac{1}{j\omega C_{in}} = \frac{1}{j\omega\left(C_{GS} + \left(1 + A\right)C_{GD}\right)}$$
(6)

in which A is the voltage gain of the amplifier, the relationship of V_{GS} and V_{in} is

$$V_{GS} = \frac{Z_{in}}{Z_g + Z_{in}} V_{in}$$
(7)

and V_{GS} is a linear and delayed version of V_{in} on linear amplification

$$V_{GS}(t) = CV_{in}(t - t_0)$$
(8)

Note that by introducing a parallel inverse nonlinear characteristic component at the input of the amplifier can reduce the distortion, which a PMOS capacitance can be a good choice to compensate the nonlinearity of NMOS input capacitance. In other words, the input impedance Z_{in} is near a constant for a wide range of V_{GS} due to the inverse characteristic of the PMOS capacitance from the NMOS counterpart. The Behavior of NMOS C_{GS} and C_{GD} in

different operation region is shown in Fig.13 (Razavi, 2000), where W is the width of the NMOS device, L is the effective length of the NMOS device. C_{OX} is the oxide capacitance per unit width, and the overlap capacitance per unit width is denoted by C_{OV}. If the device is off, $C_{GD} = C_{GS} = WC_{OV}$ and the gate-bulk capacitance comprises the series combination of the gate oxide capacitance and the depletion region capacitance.
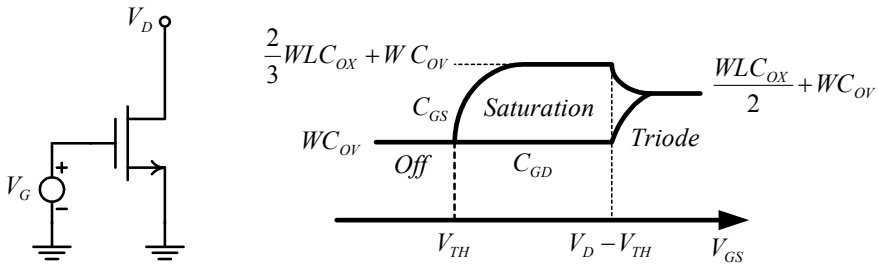

Fig. 13. Variation of gate-source and gate-drain capacitance versus V_{GS}.

If the device is operating at triode region, such that S and D have approximately equal voltages, then the gate-channel (WLC_{OX}) is divided equally and $C_{GD}=C_{GS}= (WLC_{OX})/2+WC_{OV}$. On the other hand, the gate-drain capacitance of a MOSFET is roughly equal to WC_{OV} for the saturation mode operation. The potential difference between the gate and channel varying from V_{TH} at the source to V_D-V_{TH} at the pinch-off point results in a non-uniform vertical electric field in the gate oxide along the channel. It can be proved that the gate-source capacitance equals to $(2/3)WLC_{OX}$ (Muller & Kamins, 1986). Thus, $C_{GS}=(2/3)WLC_{OX}+WC_{OV}$.

The dependence of a p-substrate MOS capacitance on voltage is shown in Fig.14 (Singh, 1994), in which V_{fb} represents flat-band voltage and V_T represents threshold voltage. In accumulation region (negative V_G), the holes accumulate at the oxide-semiconductor interface. Because holes are majority carriers, the response time is fast enough. As the gate voltage becomes positive, the interface is depleted of holes and attracts minority carriers. The depletion capacitance becomes important in this region. When the device gets more and more depleted, the value of C_{MOS} decreases to C_{MOS}(min).

At inversion condition, the depletion width reaches its maximum width. If the bias increases further, the free electrons in the p-substrate start to collect in the inversion region, whereas the depletion width remains unchanged with bias. The required excess free electrons are introduced into the channel by electron-hole generation. Since the generation process takes a certain amount of time, the inversion sheet charge can follow the bias voltage only if the voltage change speed is slow. If the variations are fast, the electron-hole generation cannot catch up the variations. The capacitance due to the free electrons has no contribution and the MOS capacitance is dominated by the original depletion capacitance. Therefore, under high-frequency conditions, the capacitance does not show a turnaround and remains at the C_{MOS}(min) as shown in Fig. 14.
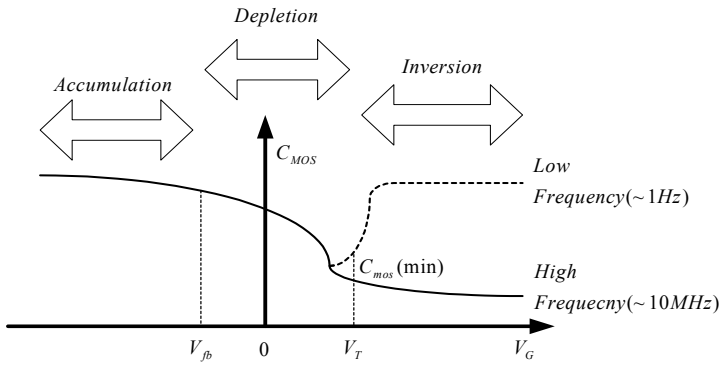
Fig. 14. Dependence of a P-substrate MOS capacitor versus voltage.

5.2 PMOS capacitance compensation technique

As shown in Fig.15, we can use this inverse capacitance characteristic to compensate the nonlinearity of NMOS input capacitance.
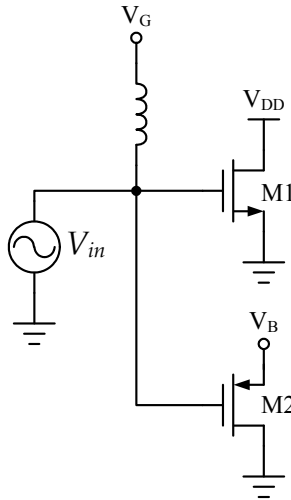
Fig. 15. Schematic of the PMOS capacitance compensation PA.

The Hspice simulation results of the NMOS and PMOS input capacitance (C_{gs} and C_{gd}) are shown in Fig.16 (a) and (b), respectively.
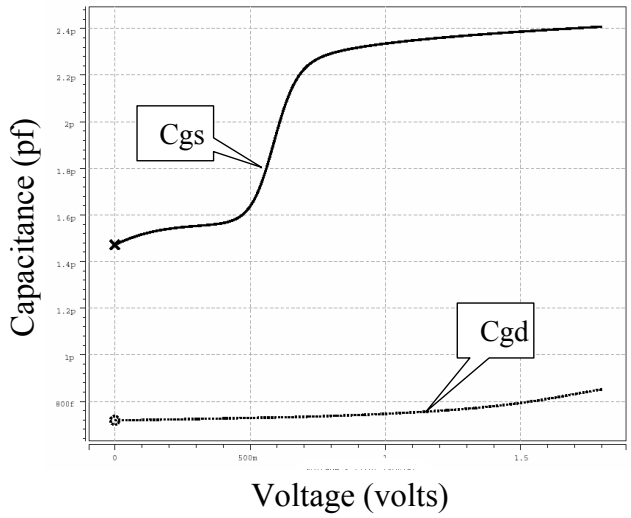
Fig. 16(a). Capacitances of C_{gs} and C_{gd} versus V_{gs} for NMOS device (W=1920μm, L=0.18μm).
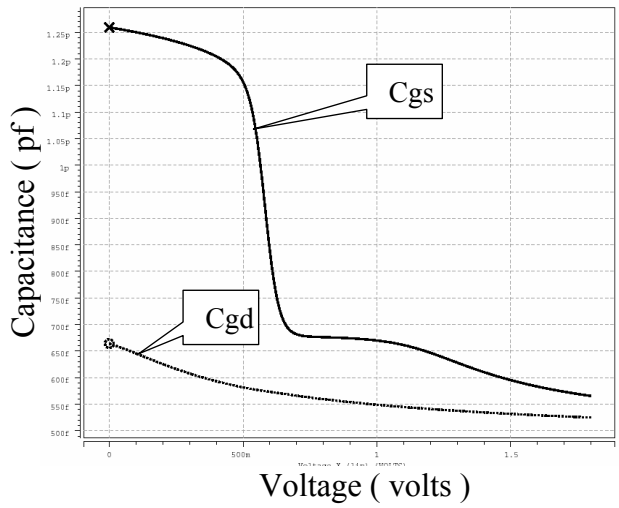
Fig. 16(b). Capacitances of C_{gs} and C_{gd} versus V_{gs} for PMOS device (W=1280μm, L=0.18μm).

The total input capacitance of the NMOS and PMOS devices is shown in Fig.17. Obviously, we can use this inverse capacitance characteristic to compensate the nonlinearity of NMOS input capacitance.
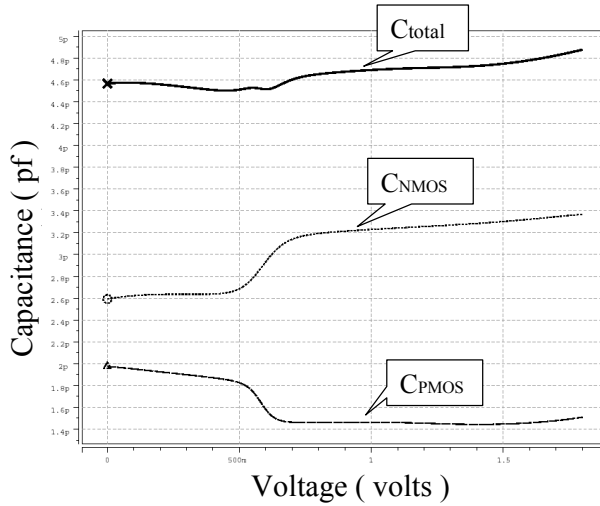
Fig. 17. Total gate input capacitance with PMOS capacitance compensation.

5.3 NMOS diode linearizer technique

The newly proposed approach is the diode linearizer which can be integrated in the PA design. The integrated diode linearizer in HBT PA can effectively improve the gain compression and phase distortion performances from the gate dc bias level (V_{GS}). Notice that the dc bias level decreases as the input power increases. A PA uses an integrated diode-connected NMOS transistor as the function of diode linearizer is shown in Fig.18. A similar technique by using nonlinear capacitance cancellation in CMOS PA designs has been reported in (Yen & Chuang, 2003).
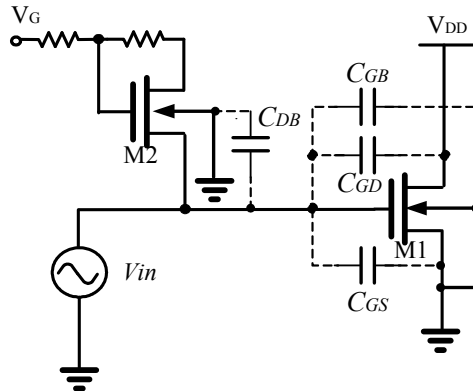
Fig. 18. Schematic of NMOS diode linearizer PA with parasitic capacitors.

For a first-order approximation, the oxide–related gate capacitances C_{GS}, C_{GD}, and C_{GB} of M1 are given by (Massobrio & Antognetti, 1993)

$$C_{GS} = \frac{2}{3} C_{OX} + C_{GS0} W \qquad (9)$$

$$C_{GD} = C_{GD0} W \qquad (10)$$

$$C_{GB} = C_{GB0} L_{eff} \qquad (11)$$

for M1 operating at the saturation mode, in which L_{eff} is the effective channel length, W is the width of the channel, C_{OX} is the gate oxide capacitance, C_{GS0}, C_{GD0}, C_{GB0} are the voltage-independent overlap capacitances per meter among the gate and the other terminals outside the channel region and

$$C_{GS} = C_{OX} \left\{ 1 - \left[\frac{V_{GS} - V_{DS} - V_{TH}}{2(V_{GS} - V_{TH}) - V_{DS}} \right]^2 \right\} + C_{GS0} W \qquad (12)$$

$$C_{GD} = C_{OX} \left\{ 1 - \left[\frac{V_{GS} - V_{TH}}{2(V_{GS} - V_{TH}) - V_{DS}} \right]^2 \right\} + C_{GD0} W \qquad (13)$$

$$C_{GB} = C_{GB0} L_{eff} \qquad (14)$$

for M1 operating at the triode mode, where V_{TH} is threshold voltage.

On the other hand, due to the depletion charge surrounding the respective drain diffusion region embedded in the substrate, the junction capacitance C_{DB} of M2 is given by

$$C_{DB} = \frac{C_j A_D}{\left(1 + V_{DB} / \varphi_j\right)^{m_j}} + \frac{C_{jsw} P_D}{\left(1 + V_{DB} / \varphi_j\right)^{m_{jsw}}} \qquad (15)$$

in which C_j and C_{jsw} are the capacitances at zero-bias voltage for square meter of area and for meter of perimeter, respectively, m_j and m_{jsw} are the substrate-junction and perimeter capacitance grading coefficients, φ_j is the junction potential, and drain-to-gate overlap capacitance C_{DG} of M2 can be described as

$$C_{DG} = C_{DG0} W \qquad (16)$$

Notice that the input-voltage-dependent capacitances C_{GS}, C_{GD} of M1 indicated in (12) and (13) increase with an increase of V_{GS} whereas the junction capacitance C_{DB} of M2 described in (15) decreases with an increase of V_{DB} ($=V_{GS}$ of M1). Therefore, with a proper choice of the dimensions of M1 and M2, a near constant total input capacitance can be achieved.

Fig. 19 shows the simulation results of the NMOS gate capacitance (C_{GS}, C_{GD}, and C_{GB}) and the NMOS diode total capacitance at drain (C_{DG} and C_{DB}). The total input capacitance of these two devices has flat curve characteristic at each V_{GS}. Clearly, it also implies the distortion due to the nonlinearity of the input capacitance can be reduced.
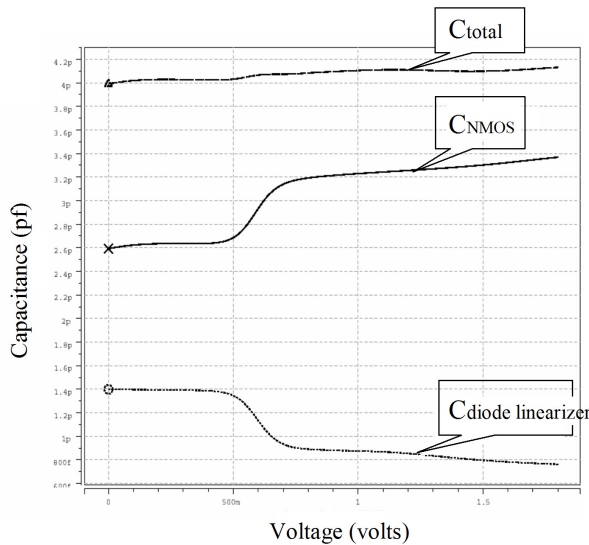
Fig. 19. Total gate input capacitance with diode linearizer for TSMC 1.8V RF MOS device.

Also, the P1dB simulation results indicate this diode-linearizer bias technique can improve 2-dB linear gain than the conventional resistance bias approach as shown in Fig.20. Note that the device dimensions of the CMOS PA can reach millimeter scale, which implies that the parasitic capacitances C_{p1} and C_{p2} can degrade the gain and power-added efficiency of the PA due to the large parasitic capacitances (Jeffrey et al., 2001).
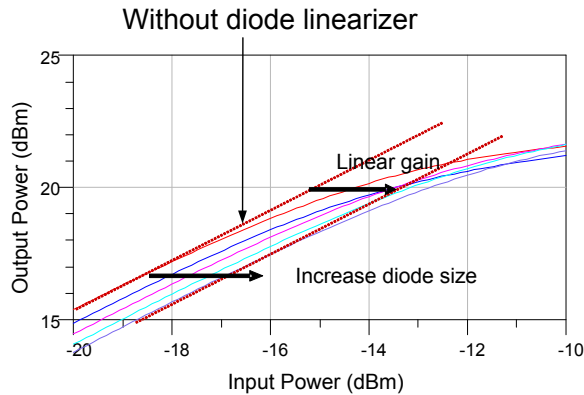
Fig. 20. P1dB simulation results of the diode-linearizer bias approach for different diode size.

5.4 Parallel inductor compensation diode technique

CMOS cascode amplifier architecture with the parallel inductor is shown in Fig. 21. Notice that the large device sizes of CMOS PA can lead to large parasitic capacitances, C_{p1} and C_{p2},

which degrade the gain and power-added-efficiency of PAs (Jeffrey et al., 2001). In order to increase the power gain of CMOS PAs and reduce the currents required to charge and discharge the parasitic capacitors at these nodes, an inductor, Ltank is used across the differential cascode nodes to produce a resonant tank at these nodes. Since the power gain increases, the 1-dB compression point is extended to a higher value and the linearity can be improved by this kind of circuit.
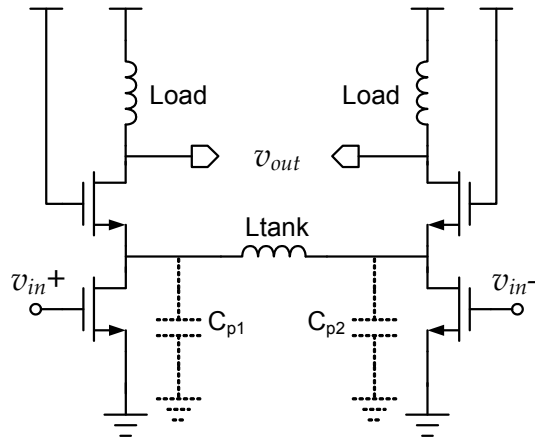
Fig. 21. CMOS cascode amplifier with a parallel inductor.

6. Self-Biased and Bootstrapped Techniques

Self-biased and bootstrapped techniques can relax the design restriction due to hot carrier degradation in power amplifiers and alleviate the requirement of using thick-oxide transistors. Note that the transistors have poor RF performance compared with the standard transistors available in the same process. Fig. 22 shows no performance degradation after ten days of continuous operation under maximum output power at 2.4-V supply voltage (Sowlati & Leenaerts, 2003); (Mertens & Steyaert, 2002).

There are two main issues in the design of power amplifiers in submicron CMOS, namely, oxide breakdown and hot carrier effect. Both of these are even worse as the technology scales. The oxide breakdown is a catastrophic effect and sets a limit on the maximum signal swing on drain. The hot carrier effect, on the other hand, is a reliability issue. It increases the threshold voltage and consequently degrades the performance of the device. The recommended voltage to avoid hot carrier degradation is usually based on dc/transient reliability tests. For production requirements, the recommended voltage is 5%–10% above the maximum allowed supply voltage to guarantee a product lifetime of ten years. For a 0.18-μm process, this leads to a maximum dc drain–gate voltage of 2 volts. CMOS power amplifiers have been reported with the dc voltage below the recommended voltage with the dc RF voltage levels of exceeding the maximum allowed value (Fallesen & Asbeck, 2001). The performance degradation due to hot carrier becomes evident during the first few hours, and the output power of the amplifier decreases in the order of 1 dB after 70–80 hours of continuous operation (Vathulya et al., 2001).

Fig. 22. Output power versus time of continuous operation.

Cascode configuration and thick-oxide transistors (Yoo & Huang, 2001; Kuo & Lusignan, 2001) have been used to eliminate the effects of oxide breakdown voltage and the hot carrier degradation, which allows the use of a larger supply voltage. So far, in cascode power amplifiers, the common-gate transistor has had a constant dc voltage with an ac (RF) ground. Under lar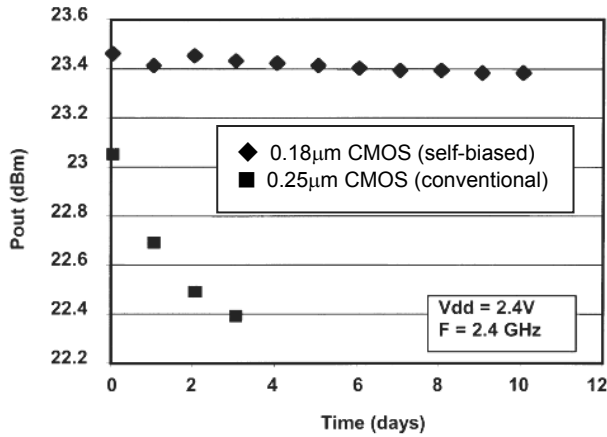ge signal operation, the voltage swing on the gate–drain of the common-gate transistor becomes larger than that of the common-source transistor. Therefore, the common-gate transistor becomes a bottleneck in terms of breakdown or hot carrier degradation. In (Yoo & Huang, 2001), the 900-MHz 0.2-µm CMOS cascode power stage uses a combination of standard and thick-oxide devices (standard device for the common-source and thick-oxide device for the common-gate). The thick-oxide device is equivalent to a device in 0.35-um process which can tolerate a much larger voltage. However, a thick-oxide device does not have the same high-frequency performance of the standard device. A cutoff frequency f_t of 26 GHz is typical for thick oxide in a 0.2-µm CMOS compared with its standard device which has a typical f_t of 50 GHz. The thick-oxide device basically provides a lower gain at RF. In a cascode combination of thick and standard devices, the thick device limits the high-frequency performance. In other words, even though we use a more advanced technology (0.18-µm process compared with 0.35-µm process), we cannot exploit the higher frequency performance of the scaled-down devices.

6.1 Conventional cascode power stage
A conventional cascode amplifier is shown in Fig. 23(a), in which transistors M1 and M2 configured as common source (CS) and common gate (CG) amplifiers, respectively. The RF signal is applied to G1. Gate G2 is RF grounded with a dc value of V_{DC} which can be equal to the supply voltage V_{DD}. The dc voltage at D2 is equal to the supply voltage with an RF voltage swing around this value. At maximum output power, the voltage at D2 swings down close to zero and up to twice V_{DD}. In order to increase the efficiency, the voltage can be shaped with the choice of the matching network. In the cascode configuration, transistor M1 has a smaller drain–gate voltage swing. This is because the voltage at D1 is always lower

than voltage at G2 by an amount equal to the gate–source voltage of G2. Consequently, the supply voltage is limited by the breakdown voltage of M2 rather than M1. This can also be observed from Fig. 23(b) which shows the time domain voltage waveforms for this amplifier. In this simulation, the supply voltage is 2.4 V and the operating frequency is 5.25 GHz.

(a)

(b)
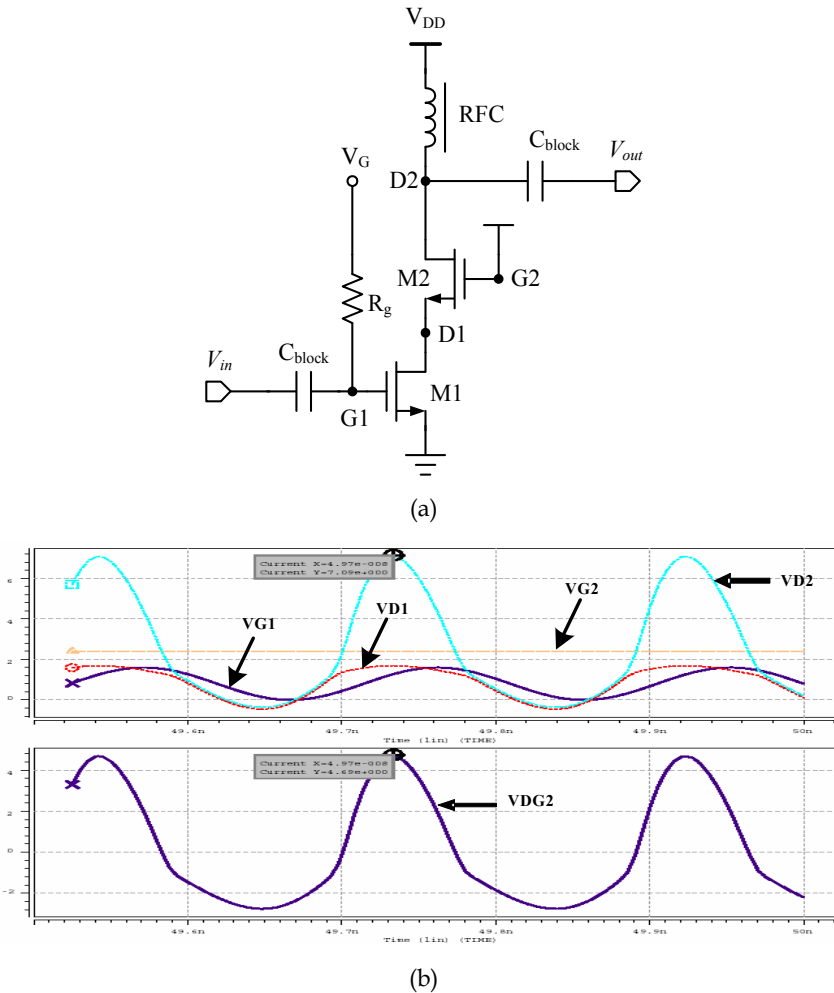
Fig. 23. (a).Conventional cascode amplifier (b).Voltage waveforms versus time for V_G=0.8V, V_{in}=0.8$sin\omega t$, and ω=5.25GHz.

6.2 Self-biased cascode power stage

To overcome the breakdown voltage limitation problem of M2, a self-biased cascode transistor is proposed as shown in Fig.24(a), which it allows RF swing at G2. This enables us to design the PA such that both transistors experience the same maximum drain–gate

voltage. Consequently, we can have a larger signal swing at D2 before encountering hot carrier degradation.

(a)

(b)
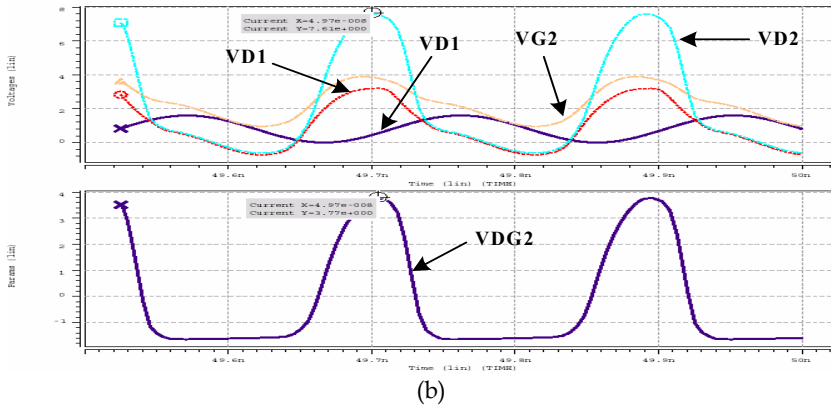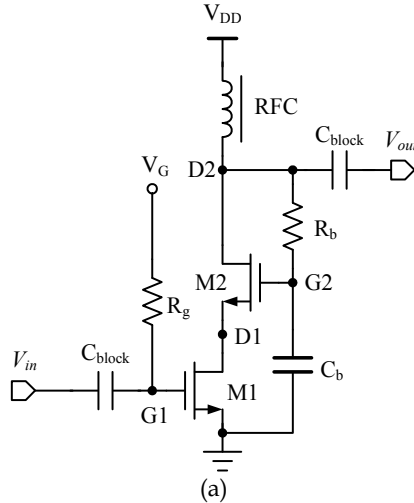
Fig. 24. Operational waveforms of (a).Self-biased cascode amplifier   (b).Voltage waveforms versus time for V_G=0.8V, V_{in}=0.8sinωt, R_b=0.75KΩ, C_b=2.4 pf,and ω=5.25GHz.

The bias for G2 is provided by R_b–C_b network, for which no extra bondpad is required. The dc voltage applied to G2 is the same as the dc voltage applied to D2. The RF swing at D2 is attenuated by the low-pass nature of R_b–C_b network as shown in Fig.24(b). The values of R_b and C_b can be chosen for optimum performance and for equal gate–drain signal swings on M1 and M2. As G2 follows the RF variation of D2 in both positive and negative swings around its dc value, a non-optimal gain performance is obtained (compared with a cascode with RF ground at G2). However, as long as both M1 and M2 go from saturation into triode under large-signal operation, the maximum output power and PAE are not degraded. The effect of the self-biased concept is demonstrated in Fig.24(b). Here, the same dimensions of

the devices are used as for the case presented in Fig. 23(b). A reduction of more than 20% in the drain–gate voltage of M2 is actually obtained.

6.3 Boot-strapped cascode power stage

To further extend this idea, we can add a resistive-diode boosting so that the positive swing of G2 can be made larger than the negative swing as shown in Fig. 25(a). By choosing the value of R_d and the size of the diode connected transistor M3, we can specify the threshold voltage at which the R_d–M3 starts conducting and boosting the positive swing at G2. This extra path enables G2 to follow the rise in D2 with a smaller attenuation than the fall in D2. During this transient response, the average charge stored on C_b increases causing R_d–M3 to conduct for a smaller percentage of the duty cycle. The average voltage at G2 increases up to the point where R_d–M3 no longer conducts. In steady state, the R_d–M3 path is off and the positive and negative swings at G2 are equal. Fig. 25(b) shows the drain and gate voltages of transistor M2 versus time for different values of R_d. The voltage swing at D2 is not affected by R_d, and the peak-to-peak swing of VG2 depends on R_b–C_b and not R_d. However, as the value of R_d is reduced, the average voltage of VG2 increases. In Class-E PA design, the voltage swing can be about three times the supply voltage (with a larger positive swing than negative around supply). In this situation, the bootstrapped cascode configuration can be employed to have the same maximum voltage swings at gate–drain of M1 and M2. Therefore, a larger supply voltage can be applied, resulting in a higher output power. For Class-AB/B design, where the signal has roughly the same positive and negative swings around the supply voltage, the self-biased cascade provides the required swing on G2.

(a)

(b)
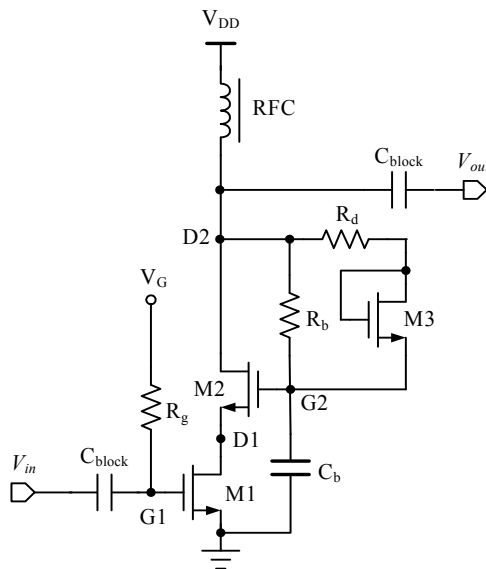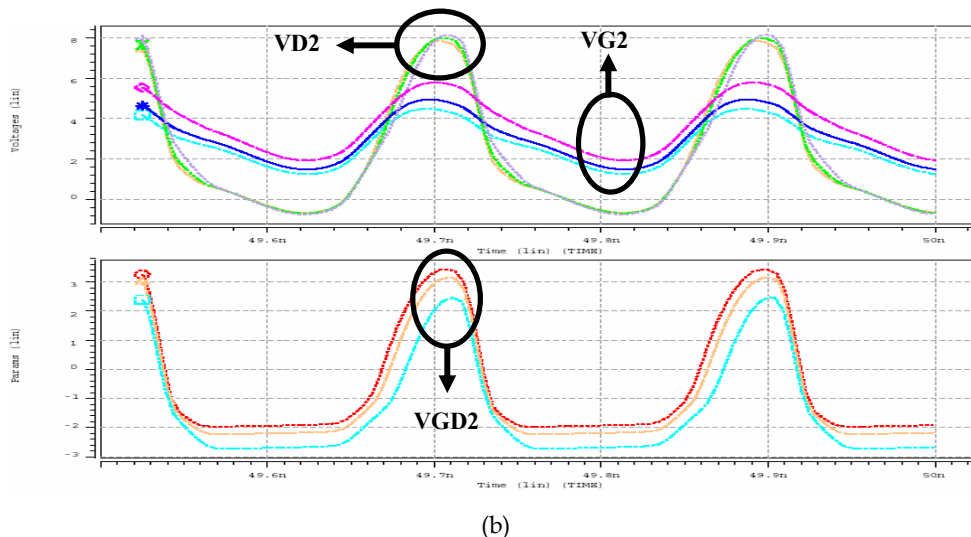
Fig. 25. (a). Bootstrapped cascode power stage. (b). Voltage waveforms.

7. Case Study–Implementation of a 5.25-GHz CMOS Cascode Power Amplifier for 802.11a WLAN

In the case study, we investigate a 5.25-GHz highly integrated CMOS class-AB power amplifier for IEEE 802.11a WLAN. The proposed power amplifier is implemented with a two gain-stage structure which is followed by an off-chip output matching circuit. Moreover, transistor-level compensation techniques are employed to improve the linearity. The power amplifier is designed with an on-chip input matching circuit while the output matching circuit translates the signal power from 50-Ω to 20-Ω load resistance. The measured results indicate over 20% power-added efficiency, over 20-dBm output power, and 28.6-dBm output IP3. All the specifications are based on 50-Ω input impedance at 2.4V supply voltage.

7.1 Introduction

Integration with a CMOS process is the key challenge for the state-of-the-art systems-on-a-chip (SOC) design approach. Conventionally, wireless local area network (WLAN) has been implemented with a multi-chip approach. However, the integration of baseband and RF front-end circuits with a CMOS process is the most promising approach to achieve highly integrated level, low power consumption, and low cost for a WLAN system. A challenging functional block in designing a wireless communication transceiver is the power amplifier due to the trade-offs between supply voltage, output power, power efficiency, and linearity, which the problem may couple with spectrum efficiency and leading to an even more difficult dilemma.

In order to achieve a higher spectrum-efficiency, the new OFDM (Orthogonal Frequency Division Multiplexing) based WLAN standards use non-constant envelope modulation,

which the linearity of the power amplifier is a key parameter as it is closely related to power consumption and distortion. Moreover, class-AB power amplifiers are widely used in wireless transceiver design due to their high efficiency and relatively high linearity. Transistor-level compensation techniques to enhance the linearity of a CMOS power amplifier are investigated in this case study. On the other hand, cascode configuration has been employed to eliminate the effects of oxide breakdown voltage and hot carrier degradation, which allows the use of a higher supply voltage. A self-biased technique with thin-oxide MOS is presented, which it can relax the restriction due to the hot carrier degradation in power amplifiers and alleviate the conventional requirement of using thick-oxide transistors.

7.2 Implementation of a two-stage cascode differential power amplifier

Some transistor-level linearization techniques have been employed in the radio-frequency PA design including nonlinear capacitance cancellation in CMOS PA design, PMOS cancellation, parallel diode with a bias feed resistance, and varactor cancellation. In this case study, a diode linearizer as presented in Section 5.3 is integrated in the proposed class-AB PA design, which it can effectively reduce the gain compression and phase distortion. In order to increase the power gain of the CMOS PA and reduce the current required to charge and discharge the parasitic capacitors of these nodes, the inductor *Ltank* is employed across the differential cascode nodes between drain and source connections, which acts as a resonant tank at these nodes as shown in Fig. 21 of Section 5.4.

The two main issues in the design of power amplifiers in deep-submicron CMOS technologies, namely, the oxide breakdown and the hot carrier effect, which become even worse as the technology scales down. The oxide breakdown is a catastrophic effect and sets a limitation of the maximum signal swing on the drain. The hot carrier effect increases the threshold voltage and consequently degrades the performance of a device. To avoid hot carrier degradation, the operating voltage is usually based on dc/transient reliability tests. For production requirements, the voltage is 5%–10% above the maximum allowed supply voltage to ensure a product lifetime. For a 0.18-μm process, this leads to a maximum dc drain-to-gate voltage of approximately 2V.

Cascode configuration and thick-oxide transistors have been employed to eliminate the effects of oxide breakdown voltage and the hot carrier degradation, which allow the use of a higher supply voltage. Under large signal operation, the voltage swing across the gate and drain nodes of the common-gate transistor becomes larger than that of the common-source transistor. Therefore, the common-gate transistor becomes the bottleneck in terms of breakdown or hot carrier degradation. Since the characteristic of thick-oxide devices is equivalent to a device of 0.35-μm process, a combination of standard and thick-oxide devices (standard device for the common-source and thick-oxide device for the common-gate) can tolerate a much higher voltage, which was demonstrated in the implementation of a 900-MHz, 0.2-μm CMOS cascode power stage.

The power amplifier in this case study operates at 5.15GHz-5.35 GHz frequency band, the maximum output power level is over 20dBm, and drain efficiency is over 20%. Moreover, the power amplifier employs a NMOS device to compensate the nonlinear input capacitance variation. By taking advantage of the NMOS device, the nonlinear capacitance can be compensated to nearly constant in the input of the common source device, which in turn improves the linearity. Miller's capacitance effect has also been alleviated by the

employment of the two-stage cascode differential architecture. Furthermore, the fully differential topology can bring the advantages of even order harmonics suppression and better immunity against noise from power supply and the lossy substrate. Note that for a 0.18-μm CMOS technology, the cut-off frequency f_t exceeds 60GHz, the minimum noise figure NF_{min} is below 0.5dB and a threshold voltage of 0.4V, which is a promising technology to implement a high-frequency PA operating at the frequency band.

On the other hand, a self-biased cascode structure presented in Section 6.2 does not necessitate thick-oxide transistors since it can have a larger signal swing at node D2 before encountering hot carrier degradation and this structure is employed in our power amplifier design to alleviate the hot carrier effect.

(a). Matching Network with Bond-Wire and Pad

The bond-wire can be modeled with the series connection of an inductor and a resistor, which the corresponding inductance and resistance are about 0.8nH/mm and 0.16Ω/nH, respectively for the diameter of 25-μm bond-wires. Moreover, constructed from a stack of metal6 (20kA in height), via5 (6kA), and metal5 (5.8kA) layers, the pad occupies an area of 80×80μm², which is equivalent to the series connection of a 625Ω-resistor and a 0.0625pF-capacitor to ground. Therefore, the bond-wire and pad can be modeled with the equivalent circuit which can be easily matched with a π-type matching network.

(b). Driver Stage

The first stage of the power amplifier is configured at class-A operation to provide the sufficient gain and linearity for the design. The schematic of the stage is shown in Fig.26, which the center frequency is determined by the LC high-pass matching network constructed with L_1, C_1, L_2, C_2, and parasitic capacitors. L_1, L_2, C_{p3}, and C_{p4} perform the inter-stage impedance matching at 5.15GHz-5.35GHz.
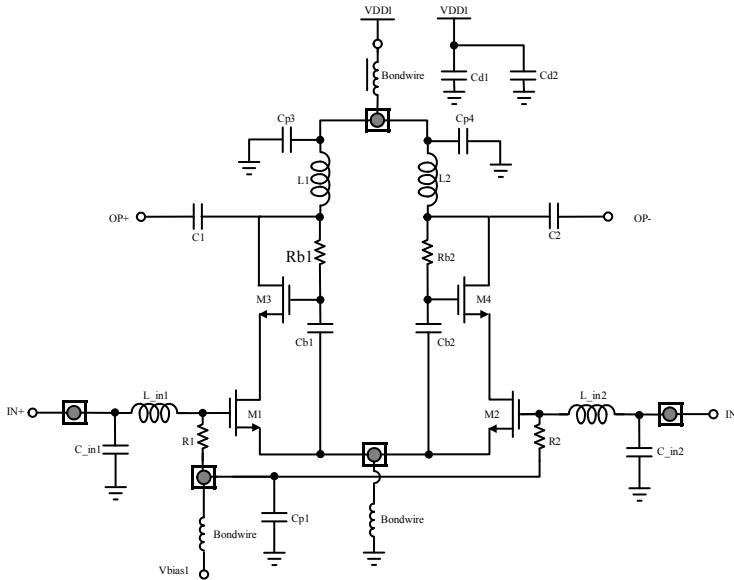
Fig. 26. Schematic of the driver stage.

(c). Power Stage

Fig. 27 shows the second stage of the NMOS diode linearizer PA which is configured at class-AB operation to obtain sufficient power efficiency and linearity. The NMOS diode linearizer is constructed with Md_1, R_3, R_4 and Md_2, R_5, R_6. The inductor L_t resonates with the parasitic capacitance of the cascode amplifier, which it can improve the intermodulation distortion (IMD) performance.
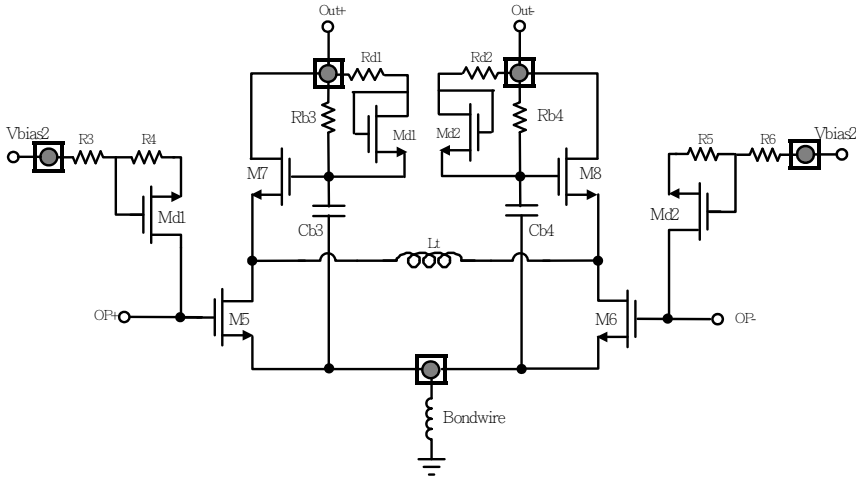
Fig. 27. Schematic of the power stage.

Due to large AC current swing of the power stage, we employ individual driver-stage and power-stage supply voltages (VDD1 and VDD2). Note that the interaction between the two supply voltages may lead to unstable operation. The pad can be modeled with a shunt connection of a resistor and a capacitor, and then series connection with an inductor. Fig.28 shows an output-matching network with an output pad.
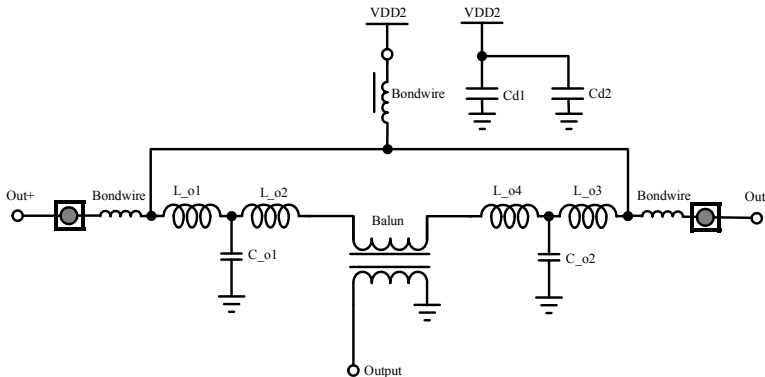
Fig. 28. Schematic of output matching network.

7.3 Experimental results

The experimental test chip is designed and fabricated with the TSMC 0.18-µm single-poly-six-metal (1P6M) salicide CMOS technology. Fig.29 shows the chip microphotograph, which the chip area is 758µm × 1741µm including the bonding pads. The active components are 1.8V NMOS devices to alleviate voltage stress. The transistors comprise the dimension of 320µm × 0.18µm MOS finger cells.
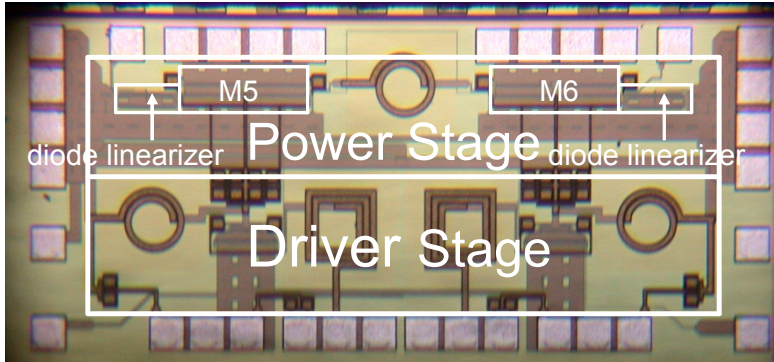
Fig. 29. Chip microphotograph of the NMOS diode compensation PA.

The driver stage of the power amplifier employs a common-source amplifier with on-chip spiral inductors while the power-stage loads of the power amplifier employ a quarter-wave length transmission line to form an RF choke on PCB to reduce the loss of the output matching networks. The number of fingers of each stage is made large to reduce the polysilicon-gate resistance since the resistance can degrade the RF signal and increase noise. The output paths are the top thick metals to avoid voltage drop and electron migration problems. Output pads are located close to the output devices in order to eliminate extra unnecessary resistance and capacitance.

The internal large AC coupled capacitors are used between power supply buses and the ground buses to form RF virtual ground. The matching and symmetry should be arranged very carefully for the differential structure. Each device cell has dummy transistors on the edges to ensure the same environment of all MOS transistors and square metal-insulator-metal (MIM) capacitors with thin oxide are employed throughout the power amplifier. Moreover, the double guard rings around the spiral inductors prevent substrate noise from other circuits.

Fig. 30 is the measurement setup for the power amplifier, in which a single-ended 5.25-GHz signal is developed by the RF signal generator. The input balun converts a single-ended signal to differential output signals, and applies the signals to the device under test (DUT). The output balun converts the differential signals to a single-ended signal and delivers the resultant signal to the input terminal of spectrum analyzer to analyze the output fundamental power signal and harmonics.

During measurement, the bare die of the CMOS PA has been directly mounted to the ground plane of the lossy PCB to minimize the length of a bonding wire, which reduces the parasitic effects of commercial packages and alleviates the thermal effect. The dielectric

constant of the PCB used in this work was a standard FR4 process with a relative dielectric constant of 4.5.
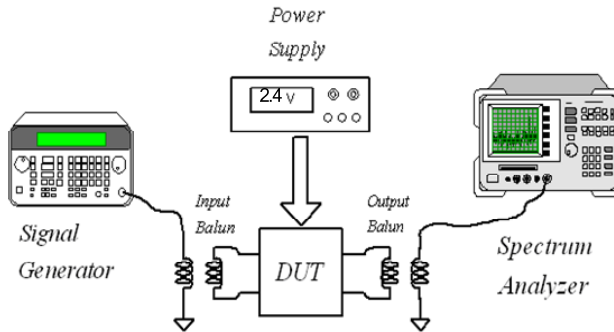
Fig. 30. Measurement setup for the power amplifier.

Because the poor quality and large current tolerance ability of the standard spiral inductor in CMOS process, the final stage load inductor excludes from the chip. This test chip employs high-frequency inductors and capacitors as the power stage load matching network to improve the efficiency of the PA and overcome the large current issue. Fig.31 shows the measured output signal spectrum of the NMOS diode compensated PA, which it demonstrates a 16.5-dBm output power when the input power is -2dBm. Fig.32 shows the output power versus input power, which the measured P1dB is approximately 16.5dBm. Note that the measured data includes the cable, balun, PCB board, and pad losses for input and output ports and each port has about 4-dB power loss.
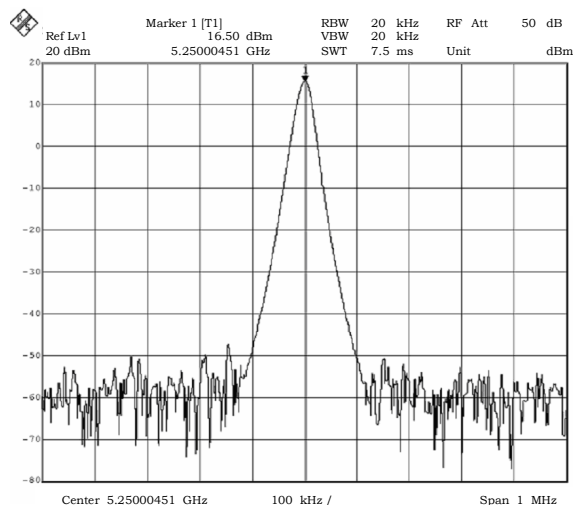
Fig. 31. Measured output power of NMOS diode compensated PA when input power is -2dBm.
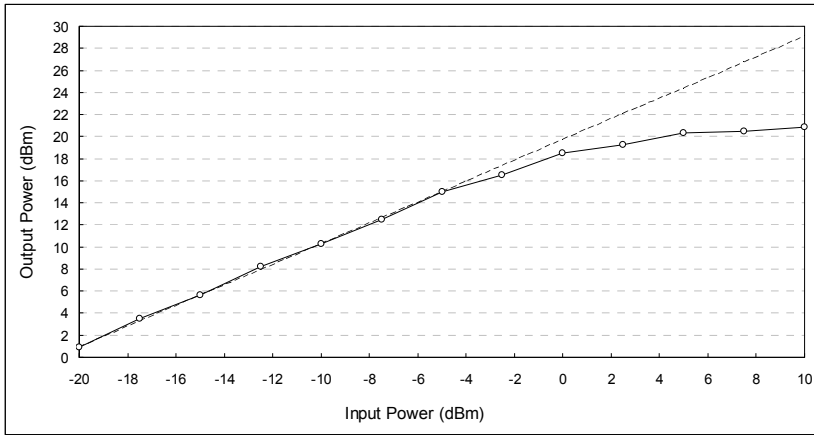
Fig. 32. Output power versus input power and P1dB.

The power-added efficiency (PAE) is indicated in Fig.33. For a lower input power, the extra power losses from the measurement setup are less important since the power amplifier itself inherently has large power loss. By contrast, for a higher input power, however these extra power losses degrade the efficiency more seriously since the power loss of the power amplifier is smaller. Obviously, the measured PAE is 20.1% when the input power is -2dBm. In order to measure the linearity of the power amplifier, an IM3 two-tone test was performed. Two tones spaced at 160MHz have been applied to the input of the power amplifier. Fig.34 shows the measured output IP3 (OIP3) which is approximately 28.6dBm.
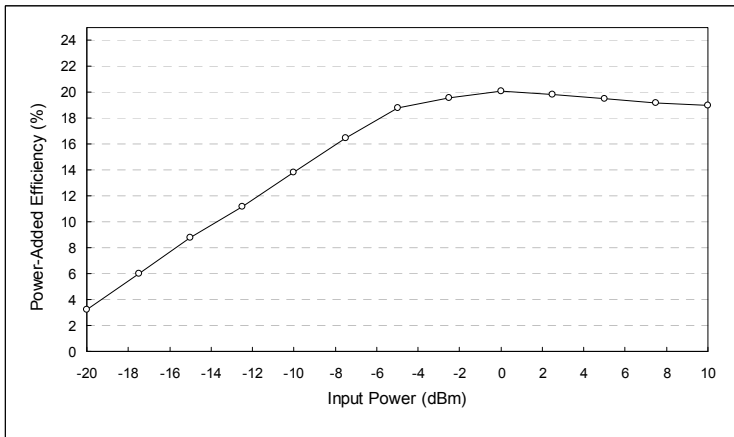
Fig. 33. Power-added efficiency versus input power.
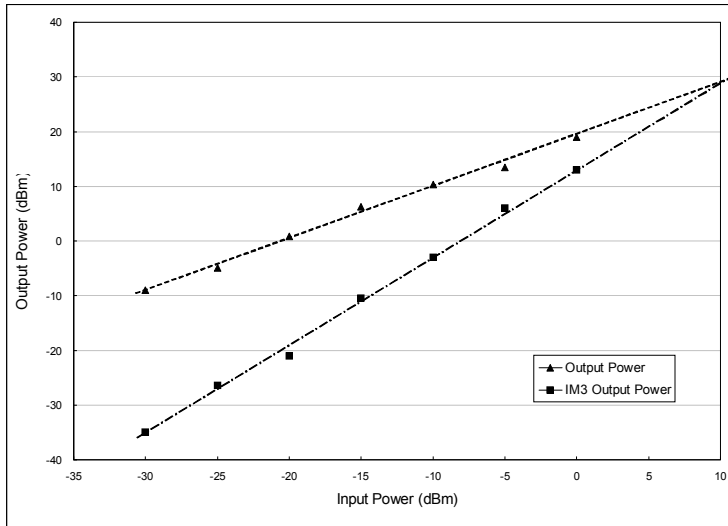
Fig. 34. Measured output IP3.

Table 1 summarizes the measured key performance feature of the power amplifier, which shows comparable performance in terms of linearity and intermodulation distortion under the measurement setup.

Technology	TSMC 0.18-μm 1P6M RF CMOS
Supply voltage	2.4V
Center frequency	5.25GHz
Maximum output power	20.9dBm
Power-added efficiency @Pout = 16 dBm	20.1%
Output P1dB	16.5dBm
Output IP3	28.6dBm
DC current of driver stage	44mA
DC current of power stage	112mA

Table 1. Measured performance summary

8. Conclusion

In this chapter, we have presented the design aspects of the class-AB linear power amplifier. The proposition of the linear power amplifier for high spectrum-efficiency communications in CMOS process technology is mainly due to the integration of a single-chip RF radio. The inherently theoretical high-power efficiency characteristic is especially suitable for wireless communication applications. Moreover, linearization enhancement techniques have also been investigated, which makes the power amplifier be practically employed in high spectrum-efficiency communications.

Finally, in the case study a 5.25-GHz, high-linearity, class-AB power amplifier has been investigated and integrated on a chip in 0.18-μm RF CMOS technology. The CMOS PA uses a NMOS diode to compensate the distortion of the PA. Requirements of the specification have been discussed and translated into circuit designs and simulation results. Experimental results indicate a good agreement with the compensation approach.

9. References

Asbeck, P. & Fallesen, C. (2000). A Polar System for RF Power Amplifiers, *The 7th International Conf. on Electronics, Circuits and Systems*, Vol. 1, pp.478-481, 2000.

Cripps, S. C. (2002). Feedback Techniques, In: *Advanced Techniques in RF Power Amplifier Design,* Norwood, MA: Artech House.

Eberle, W., et al. (2001). Digital 72Mbps 64-QAM OFDM transceiver for 5GHz wireless LAN in 0.18μm CMOS, *IEEE ISSCC Dig. Tech. Papers*, pp. 336–337, Feb. 2001.

Fallesen, C. & Asbeck, P. (2001). A 1-W 0.35-_m CMOS power amplifier for GSM-1800 with 45% PAE, *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, pp. 158–159, Feb. 2001.

Hau, G., Bishimura, T. B. & Iwata, N. (1999). 57% Efficiency, Wide Dynamic Range Linearized Heterojunction FET-Based Power Amplifier for Wide-Band CDMA Handsets, *21st Annual of GaAs IC Sym.*, pp. 295-298, 1999.

Heo, D., Gebara, E., Chen, Yoo, S., Hamai, M., Suh, Y. & Laskar, J. (2000). An Improved Deep Submicrometer MOSFET RF Nonlinear Model with New Breakdown Current Model and Drain-to-Substrate Nonlinear Coupling, *IEEE Trans. Microwave Theory Tech.*, Vol. 48, No. 12, Dec. 2000, pp. 2361-2369.

Jeffrey, A., Weldon, R., Narayanaswami, S., Rudell, J. C., Lin, L., Otsuka, M., Dedieu, S., Tee, L., Tsai, K., Lee, C. & Gray, P. R. (2001). A 1.75GHz Highly Integrated Narrow-Band CMOS Transmitter With Harmonic-Rejection Mixers, *IEEE Journal of Solid-State Circuits*, Vol. 36, No. 12, Dec. 2001, pp. 2003-2015.

Jeon, K., Kwon, Y., & Hong, S. (1997). Input Harmonics control using non-linear capacitor in GaAs FET Power Amplifier, *IEEE MTT-S Dig.*, Vol. 2, pp. 817-820, 1997.

Jeon, M., Kim, J., Kang, H., Jung, S., Lee, J. & Kwon, Y. (2002). A New 'Active' Predistortor With High Gain Using Cascode-FET Structures, *IEEE RFIC Symp.*, pp.253-256, 2002.

Johansson, M. & Mattsson, T. (1991). Transmitter Linearization Using Cartesian Feedback for Linear TDMA Modulation, *Proc. IEEE Veh. Tech. Conf.*, pp.439-444, 1991.

Kuo, T. & Lusignan, B. (2001). A 1.5-W class-F RF power amplifier in 0.25-_m CMOS technology, *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, pp. 154–155, Feb. 2001.

Massobrio, G. & Antognetti, P. (1993). *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York.

Mertens, K. L. R. & Steyaert, M. S. J. (2002). A 700-MHz 1-W fully differential CMOS class-E power amplifier, *IEEE Journal of Solid-State Circuits*, Vol.37, Feb. 2002, pp.137-141.

Morris, K. A. & McGeehan, J. P. (2000). Gain and phase matching requirements of cubic predistortion systems, *IEE Electronics Letters*, Vol.36, No. 21, Oct. 2000, pp.1822-1824.

Muller, R. S. & Kamins, T. I. (1986). *Device Electronics for Integrated Circuits*, Second Ed., New York: Wiley.

Peter, V. (1983). Reduction of Spurious Emission from Radio Transmitters by Means of Modulation Feedback, *IEE Conf. on Radio Spectrum Conservation Tech.*, pp.44-49, 1983.

Razavi, B.(1999). RF Transmitter Architectures and Circuits, *IEEE Custom Integrated Circuits Conference,* 1999.

Razavi, B. (2000). Basic MOS Device Physics, In: *Design of Analog CMOS Integrated Circuits*, McGraw-Hill.

Ryan, P. et al.(2001). A single chip PHY COFDM modem for IEEE 802.11a with integrated ADC's and DACs, *ISSCC Dig. Tech. Papers*, pp. 338–339, Feb. 2001.

Shi, B. And Sundstrom, L. (1999). Design and Implementation of A CMOS Power Feedback Linearization IC for RF Power Amplifiers, *Proc. Int. Symp. on Circuits and Systems*, Vol. 2, pp. 252-255, 1999.

Singh, J. (1994). FIELD EFFECT TRANSISTORS: MOSFET, In: *Semiconductor Devices An Introduction*, McGraw-Hill.

Sowlati, T. & Leenaerts, D. M. W. (2003). A 2.4-GHz 0.18-um CMOS Self-Biased Cascode Power Amplifier, *IEEE Journal of Solid-State Circuits*, Vol. 38, No. 8, Aug. 2003, pp. 1318-1324.

Su, D. and McFarland, W. (1997). A 2.5-V, 1-W Monolithic CMOS RF Power Amplifier, *IEEE Custom IC Conf.*, pp.189-192, 1997.

Su, D. K. & McFarland, W. J. (1998). An IC for Linearizing RF Power Amplifiers Using Envelope Elimination and Restoration, *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 12, Dec. 1998, pp. 2252-2258.

Tanaka, S., Behbahani, F. & Abidi, A. A. (1997). A Linearization Technique for CMOS RF Power Amplifiers, *Symp. VLSI Circuits Dig.*, pp.93-94, 1997.

Thomson, J. *et al.* (2002). An integrated 802.11a baseband and MAC processor, *IEEE ISSCC Dig. Tech. Papers*, 2002, pp. 126-127, Feb. 2002.

Tsai, K. and Gray, P. R. (1999). A 1.9-GHz, 1-W CMOS Class-E Power Amplifier for Wireless Communications, *IEEE Journal of Solid-State Circuits*, Vol. 34, No. 7, July 1999, pp. 962-970.

Vathulya, V., Sowlati, T. & Leenaerts, D. M. W. (2001). Class-1 Bluetooth power amplifier with 24-dBm output power and 48% PAE at 2.4 GHz in 0.25-μm CMOS, *Proc. Eur. Solid-State Circuits Conf.*, pp. 84–87, Sep. 2001.

Wang, C., Larson, L. E. & Asbeck, P. M. (2001). A Nonlinear Capacitance Cancellation Technique and its Application to a CMOS Class AB Power Amplifier, *IEEE RFIC Symp.*, pp. 39-42, 2001.

Wang, W.; Zhang, Y.P. (2004). 0.18-um CMOS Push-Pull Power Amplifier With Antenna in IC Package, IEEE Microwave and Guided Wave Letters, Vol. 14 , No. 1, Jan. 2004, pp. 13-15.

Westesson, E. & Sundstrom, L. (1999). A Complex Polynomial Predistorter Chip in CMOS For Baseband on IF Linearization of RF Power Amplifiers, *Proc. Int. Sym. on Circuits and Systems*, Vol. 1, pp. 206-209, 1999.

Woerlee, P. H., Knitel, M. F., Langevelde, R. V., Klaassen, D. B. M., Tiemeijer, L. F., Scholten, A. J. & Duijnhoven, A. T. Z. (2001). RF-CMOS Performance Trends, *IEEE Trans. on Electron Devices*, Vol. 48, No. 8, Aug. 2001, pp. 1776-1782.

Wright, A. S. & Durtler, W. G. (1992). Experimental Performance of an Adaptive Digital Linearized Power Amplifier, *IEEE Trans. Vehicular Tech.,* Vol. 41, No. 4, Nov. 1992, pp.395-400.

Yamauchi, K., Mori, K., Nakayama, M., Mitsui, Y. & Takagi, T. (1997). A Microwave Miniaturized Linearizer Using a Parallel Diode with a Bias Feed Resistance, *IEEE Trans. Microwave Theory Tech.*, Vol. 45, No. 12, Dec. 1997, pp. 2431-2434.

Yen, C. & Chuang, H. (2003). A 0.25-/spl mu/m 20-dBm 2.4-GHz CMOS power amplifier with an integrated diode linearizer, *IEEE Microwave and Guided Wave Letters*, Vol. 13, No. 2 , Feb. 2003, pp. 45–47.

Yoo, C. and Huang, Q. (2001). A Common-Gate Switched 0.9-W Class-E Power Amplifier with 41% PAE in 0.25-um CMOS, *IEEE Journal of Solid-State Circuits*, Vol. 36, No. 5, May 2001, pp. 823-830.

Yu, C., Chan, W. & Chan, W. (2000). Linearised 2GHz Amplifier for IMT-2000, *Vehicular Tech. Conf. Proc.*, Vol. 1, pp. 245-248, 2000.

Zargari, M., Su, D. K., Yue, P., Rabii, S., Weber, D., Kaczynski, B. J., Mehta, S. S., Singh, K., Mendis, S. and Wooley, B. A. (2002). A 5-GHz CMOS Transceiver for IEEE 802.11a Wireless LAN Systems, *IEEE Journal of Solid-State Circuits*, Vol. 37, No. 12, Dec. 2002, pp. 1688-1694.

Terrestrial Free-Space Optical Communications

Ghassemlooy, Z. and Popoola, W. O.
Optical Communications Research Group, NCRLab,
Northumbria University, Newcastle upon Tyne, UK

1. Introduction

Free-space optical communication (FSO) or better still laser communication is an age long technology that entails the transmission of information laden optical radiation through the atmosphere from one point to the other. The earliest form of FSO could be said to be the Alexander Graham Bell's Photophone of 1880. In his experiment, Bell modulated the Sun radiation with voice signal and transmitted it over a distance of about 200 metres. The receiver was made of a parabolic mirror with a selenium cell at its focal point. However, the experiment did not go very well because of the crudity of the devices used and the intermittent nature of the Sun radiation. The fortune of FSO changed in the 1960s with the discovery of optical sources, most importantly the laser. A flurry of FSO demonstrations was recorded in the early 1960s into 1970s. Some of these included the: spectacular transmission of television signal over a 30 mile (48 km) distance using GaAs light emitting diode by researchers working in the MIT Lincolns Laboratory in 1962, a record 118 miles (190km) transmission of voice modulated He-Ne laser between Panamint Ridge and San Gabriel Mountain, USA in May 1963 and the first TV-over-laser demonstration in March 1963 by a group of researchers working in the North American Aviation. The first laser link to handle commercial traffic was built in Japan by Nippon Electric Company (NEC) around 1970. The link was a full duplex 0.6328 μm He-Ne laser FSO between Yokohama and Tamagawa, a distance of 14 km (Goodwin, 1970).

From this time on, FSO has continued to be researched and used chiefly by the military for covert communications. FSO has also been heavily researched for deep space applications by NASA and ESA with programmes such as the then Mars Laser Communication Demonstration (MLCD) and the Semiconductor-laser Inter-satellite Link Experiment (SILEX) respectively. Although, deep space FSO lies outside the scope of our discussion here, it is worth mentioning that over the past decade, near Earth FSO were successfully demonstrated in space between satellites at data rates of up to 10 Gbps (Hemmati, 2006). In spite of early knowledge of the necessary techniques to build an operational laser communication system, the usefulness and practicality of a laser communication system was until recently questionable for many reasons (Goodwin, 1970): First, existing communications systems were adequate to handle the demands of the time. Second, considerable research and development were required to improve the reliability of components to assure reliable system operation. Third, a system in the atmosphere would

always be subject to interruption in the presence of heavy fog. Fourth, use of the system in space where atmospheric effects could be neglected required accurate pointing and tracking optical systems which were not then available. In view of these problems, it is not surprising that until now, FSO had to endure a slow penetration into the access network.

But with the rapid development and maturity of optoelectronic devices, FSO has now witnessed a re-birth. Also, the increasing demand for more bandwidth in the face of new and emerging applications implies that the old practice of relying on just one access technology to connect with the end users has to give way. These forces coupled with the recorded success of FSO in military applications have rejuvenated interest in its civil applications within the access network. Several successful field trials have been recorded in the last few years in various parts of the world which have further encouraged investments in the field. This has now culminated into the increased commercialisation and the deployment of FSO in today's communication infrastructures.

FSO has now emerged as a commercially viable alternative to radio frequency (RF) and millimetre wave wireless systems for reliable and rapid deployment of data and voice networks. RF and millimetre wave technologies wireless networks can offer data rates from tens of Mbps (point-to-multipoint) up to several hundred Mbps (point-to-point). However, there is a limitation to their market penetration due to spectrum congestion, licensing issues and interference from unlicensed bands. The future emerging license-free bands are promising, but still have certain bandwidth and range limitations compared to the FSO. The short-range FSO links are used as an alternative to the RF links for the last or first mile to provide broadband access network to businesses as well as a high bandwidth bridge between the local area networks (LANs), metropolitan area networks (MANs) and wide area networks (WANs) (Pelton, 1998).

Full duplex FSO systems running at up to 1.25 Gbps between two static nodes and covering a range of over 4 km in clear weather conditions are now common sights in today's market. Integrated FSO/fibre communication systems and wavelength division multiplexed (WDM) FSO systems are currently at experimental stages and not yet deployed in the market. One of such demonstrations is the single-mode fibre integrated 10 Gbps WDM FSO carried out in Japan (Kazaura et al., 2007). The earlier scepticism about FSO's efficacy, its dwindling acceptability by service providers and slow market penetration that bedevilled it in the 1980s are now rapidly fading away judging by the number of service providers, organisations, government and private establishments that now incorporate FSO into their network infrastructure. Terrestrial FSO has now proven to be a viable complementary technology in addressing the contemporary communication challenges; most especially the bandwidth/high data rate requirements of end users at an affordable cost. The fact that FSO is transparent to traffic type and data protocol makes its integration into the existing access network far more rapid. Nonetheless, the atmospheric channel effects such as thick fog, smoke and turbulence as well as the attainment of 99.999% availability still pose the greatest challenges to long range terrestrial FSO. One practical solution is the deployment of a hybrid FSO/RF link, where an RF link acts as a backup to the FSO.

2. Fundamentals of FSO

FSO in basic terms is the transfer of signals/data/information between two points using optical radiation as the carrier signal through an unguided channel. The data to be transported could be modulated on the intensity, phase or frequency of the optical carrier. An FSO link is essentially based on line-of sight (LOS). Thus, both the transmitter and the receiver must directly 'see' one another without any obstruction in their path for the communication link to be established. The unguided channels could be any or a combination of the space, sea-water, or the atmosphere. The emphasis here is on terrestrial FSO and as such only the atmospheric channel will be considered.

An FSO communication system can be implemented in two variants. The conventional FSO shown in Fig. 1 is for point-to-point communication with two similar transceivers; one at each end of the link. This allows for a full-duplex communication. The second variant uses the modulated retro-reflector (MRR). Laser communication links with MRRs are composed of two different terminals and hence are asymmetric links. On one end of the link, there is the MRR while the other hosts the interrogator as shown in Fig. 2. The interrogator projects a continuous wave (CW) laser beam out to the retro-reflector. The modulated retro-reflector modulates the CW beam with the input data stream. The beam is then retro-reflected back to the interrogator. The interrogator receiver collects the return beam and recovers the data stream from it. The implementation just described permits only simplex communication. A two-way communication can also be achieved with the MRR by adding a photodetector to the MRR terminal and the interrogator beam shared in a half-duplex manner. Unless otherwise stated however, the conventional FSO link is assumed throughout this chapter.
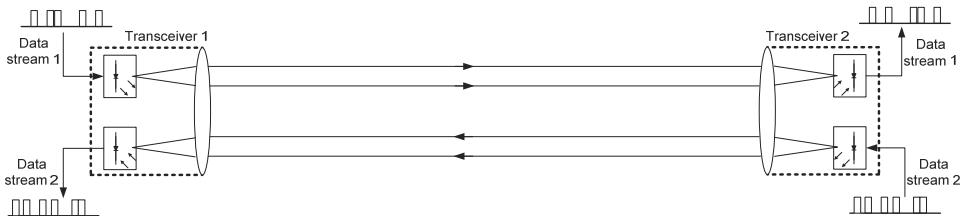
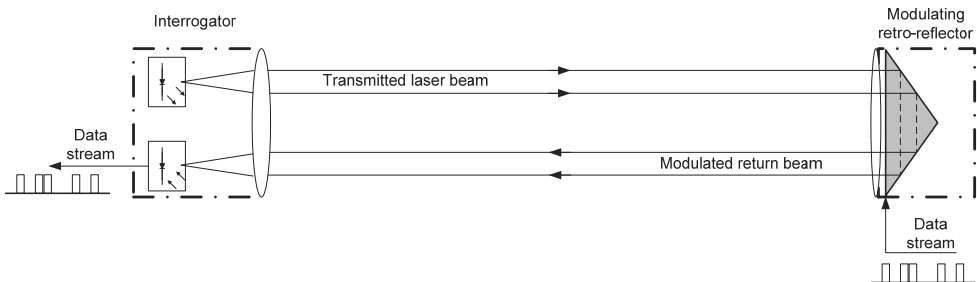Fig. 1. Conventional FOS system block diagram

Fig. 2. Modulated retro-reflector based FSO system block diagram

The basic features of FSO, areas of application and the description of each fundamental block are further discussed in the following sections.

2.1 Features of FSO
The basic features of the FSO technology are given below:

a)   *Huge modulation bandwidth* - In general, the optical carrier frequency which includes infrared, visible and ultra violet frequencies are far greater than RF. And in any communication system, the amount of data transported is directly related to the bandwidth of the modulated carrier. The allowable data bandwidth can be up to 20 % of the carrier frequency. Using optical carrier whose frequency ranges from 10^{12} – 10^{16} Hz could hence permit up to 2000 THz data bandwidth. Optical communication therefore, guarantees an increased information capacity. The usable frequency bandwidth in RF range is comparatively lower by a factor of 10^5.

b)   *Narrow beam size* - The optical radiation prides itself with an extremely narrow beam, a typical laser beam has a diffraction limit divergence of between 0.01 – 0.1 mrad (Killinger, 2002). This implies that the transmitted power is only concentrated within a very narrow area. Thus providing FSO link with adequate spatial isolation from its potential interferers. The tight spatial confinement also allows for the laser beams to operate nearly independently, providing virtually unlimited degrees of frequency reuse in many environments and makes data interception by unintended users difficult. Conversely, the narrowness of the beam implies a tighter alignment requirement.

c)   *Unlicensed spectrum* - Due to the congestion of the RF spectrum, interference from adjacent carriers is a major problem facing wireless RF communication. To minimise this interference, regulatory authorities put stringent regulations in place. To be allocated a slice of the RF spectrum therefore requires a huge fee and several months of bureaucracy. But the optical frequencies are free from all of this, at least for now. The initial set-up cost and the deployment time are then reduced and the return on investments begins to trickle in far more quickly.

d)   *Cheap* - The cost of deploying FSO is lower than that of an RF with a comparable data rate. FSO can deliver the same bandwidth as optical fibre but without the extra cost of right of way and trenching. Based on a recent finding done by 'fSONA', an FSO company based in Canada, the cost per Mbps per month based on FSO is about half that of RF based systems (Rockwell and Mecherle, 2001).

e)   *Quick to deploy and redeploy* - The time it takes for an FSO link to become fully operational starting from installation down to link alignment could be as low as four hours. The key requirement is the establishment of an unimpeded line of sight between the transmitter and the receiver. It can as well be taken down and redeployed to another location quite easily.

f)   *Weather dependent* - The performance of terrestrial FSO is tied to the atmospheric conditions. The unfixed properties of the FSO channel undoubtedly pose the greatest challenge. Although this is not peculiar to FSO as RF and satellite communication links also experience link outages during heavy rainfall and in stormy weather.

In addition to the above points, other secondary features of FSO include:

- It benefits from existing fibre optics communications optoelectronics
- It is free from and does not cause electromagnetic interference
- Unlike wired systems, FSO is a non-fixed recoverable asset
- The radiation must be within the stipulated safety limits
- Light weight and compactness
- Low power consumption
- Requires line of sight and strict alignment as a result of its beam narrowness.

2.2 Areas of application

The characteristic features of FSO discussed above make it very attractive for various applications within the access and the metro networks. It can conveniently complement other technologies (such as wired and wireless radio frequency communications, fibre-to-the-X technologies and hybrid fibre coaxial among others) in making the huge bandwidth that resides in the optical fibre backbone available to the end users. Most end users are within a short distance from the backbone – one mile or less; this makes FSO very attractive as a data bridge between the backbone and the end-users. Among other emerging areas of application, terrestrial FSO has been found suitable for use in the following areas:

a)   *Last mile access* - FSO can be used to bridge the bandwidth gap (last mile bottleneck) that exists between the end-users and the fibre optics backbone. Links ranging from 50 m up to a few km are readily available in the market with data rates covering 1 Mbps to 2.5 Gbps (Willebrand and Ghuman, 2002).

b)   *Optical fibre back up link* – Used to provide back-up against loss of data or communication breakdown in the event of damage or unavailable of the main optical fibre link.

c)   *Cellular communication back-haul* – Can be used to back-haul traffics between base stations and switching centres in the 3rd/4th generation (3G/4G) networks, as well as transporting IS-95 code division multiple access (CDMA) signals from macro-and microcell sites to the base stations.

d)   *Disaster recovery/Temporary links* – The technology finds application where a temporary link is needed be it for a conference or ad-hoc connectivity in the event of a collapse of an existing communication network.

e)   *Multi-campus communication network* – Can be used to interconnect campus networks

f)   *Difficult terrains* – For example across a river, very busy street, rail tracks or where right of way is not available or too expensive to pursue, FSO is an attractive data bridge in such instances.

3. FSO Block Diagram

The block diagram of a typical terrestrial FSO link is shown in Fig. 3. Like any other communication technologies, the FSO essentially comprises of three parts: the transmitter,

the channel and the receiver. These basic parts are further discussed in the sections that follow.
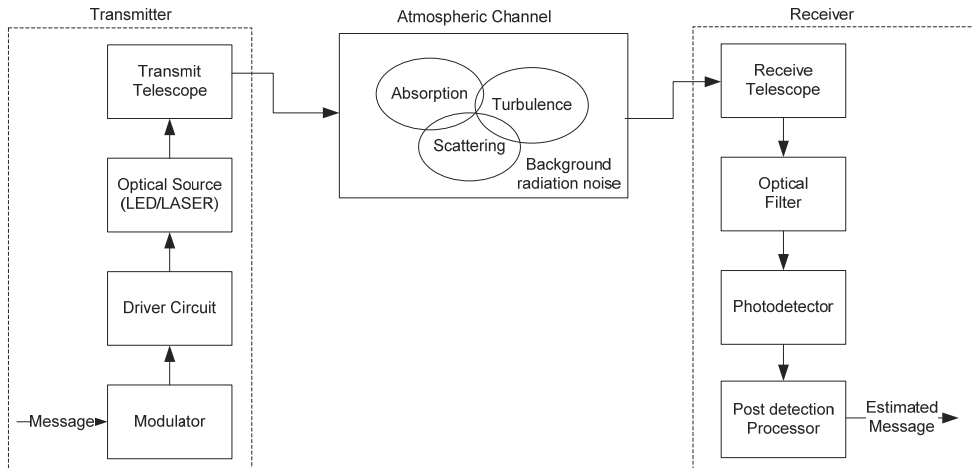
Fig. 3. Block diagram of a terrestrial FSO link

3.1 The transmitter

This functional element has the primary duty of modulating the source data onto the optical carrier which is then propagated through the atmosphere to the receiver. The most widely used modulation type is the intensity modulation (IM) in which the source data is modulated on the irradiance/intensity of the optical radiation. This is achieved by varying the driving current of the optical source directly in sympathy with the data to be transmitted or via an external modulator such as the symmetric Mach-Zehnder (SMZ) interferometer. The use of an external modulator guarantees a higher data rate than what is obtainable with direct modulation but an external modulator has a nonlinear response. Other properties of the radiated optical field such as its phase, frequency and state of polarisation can also be modulated with data/information through the use of an external modulator. The transmitter telescope collects, collimates and directs the optical radiation towards the receiver telescope at the other end of the channel. Table 1 presents a summary of commonly used sources in FSO systems.

Wavelength (nm)	Type	Remark
~850	Vertical cavity surface emitting laser	Cheap and readily available (CD lasers) No active cooling Lower power density Reliable up to ~10Gbps
~1300/~1550	Fabry-Perot Distributed-feedback lasers	Long life Lower eye safety criteria 50 times higher power density (100 mW/cm^2) Compatible with EDFA High speed, up to 40 Gbps A slope efficiency of 0.03-0.2 W/A
~10,000	Quantum cascade laser	Expensive and relative new Very fast and highly sensitive Better fog transmission characteristics. Components not readily available No penetration through glass
Near Infrared	LED	Cheaper Simpler driver circuit Lower power and lower data rates

Table 1. Optical sources

Within the 700–10,000 nm wavelength band there are a number transmission windows that are almost transparent with an attenuation of <0.2 dB/km. The majority of FSO systems are designed to operate in the 780–850 nm and 1520–1600 nm spectral windows. 780 nm - 850 nm is the most widely used because devices and components are readily available in this wavelength range and at low cost. The 1550 nm band is attractive for a number of reasons i) compatibility with the 3rd window wavelength-division multiplexing networks, ii) eye safety (about 50 times more power can be transmitted at 1550 nm than at 850 nm), and iii) reduced solar background and scattering in light haze/fog. Consequently, at 1550 nm a significantly more power can be transmitted to overcome attenuation by fog. However, the drawbacks of the 1550 nm band are slightly reduced detector sensitivity, higher component cost and a stricter alignment requirement.

3.2 The receiver
The receiver helps recover the transmitted data from the incident optical field. The receiver is composed of:

a) *The receiver telescope* - collects and focuses the incoming optical radiation on to the photodetector. It is should be noted that a large receiver telescope aperture is desirable as it collects multiple uncorrelated radiations and focuses their average on the photodetector. This is referred to as aperture averaging but a wide aperture also means more background radiation/noise,

b) *An optical band* - pass filter to reduce the amount of background radiations,

c) *A photodetector* - PIN or APD that converts the incident optical field into an electrical signal. The commonly used photodetector for in the contemporary laser

communication systems are summarised in Table 2. Germanium only detectors are generally not used in FSO because of their high dark current.

d) *Post-detection processor/decision circuit* - where the necessary amplification, filtering and signal processing necessary to guarantee a high fidelity data recovery are carried out.

Due to detector capacitance effect, higher speed detectors are inherently smaller in size (70 μm and 30 μm for 2.5 Gbps and 10 Gbps, respectively) with a limited field-of-view (FOV) that require accurate alignment. FOV of the receiver is the ratio of the detector size to the focal length (Jeganathan and Ionov): $FOV = d/f = dF\#/D$; where d is the detector diameter, f is the effective focal length, and D is the receiver aperture. The quantity $F\#$ is the f-number. For a 75 μm size detector, with $F\# = 1$ and $D = 150$ mm telescope, the FOV = ~0.5 mrad.

Material/Structure	Wavelength (nm)	Responsivity	Typical Sensitivity	Gain
Silicon PIN	300 – 1100	0.5	-34dBm@155Mbps	1
Silicon PIN, with Transimpedance amplifier	300 – 1100	0.5	-26dBm@1.25 Gbps	1
InGaAs PIN	1000 – 1700	0.9	-46dBm@155Mbps	1
Silicon APD	400 – 1000	77	-52dBm@155Mbps	150
InGaAs APD	1000 – 1700	9	-33dBm @ 1.25 Gbps	10
Quantum–well and Quatum-dot detectors	~10,000			

Table 2. FSO Photodetectors

The receiver detection process can be classified into:

a) *Direct detection receiver* - This type of receiver detects the instantaneous intensity or power of the optical radiation impinging on the photodetector. Hence, the output of the photodetector is proportional to the power of the incident field. Its implementation is very simple and most suitable for intensity modulated optical systems (Gagliardi and Karp, 1995, Pratt, 1969). The block diagram of direct detection receiver is shown in Fig. 4.
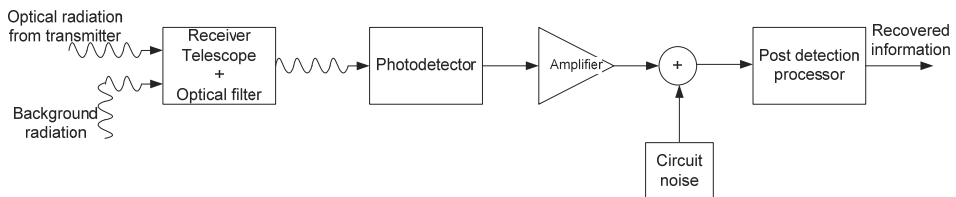

Fig. 4. The block diagram of a direct detection optical receiver.

b)   *Coherent detection receiver* – The coherent receiver whose block diagram is shown in Fig. 5 works based on the photo-mixing phenomenon. The incoming optical field is mixed with another locally generated optical field on the surface of the photodetector. The coherent receiver can be further divided into homodyne and heterodyne receivers. In homodyne receivers, the frequency/wavelength of the local (optical) oscillator is exactly the same as that of the incoming radiation while in heterodyne detection, the incoming radiation and the local oscillator frequencies are different. In contrast to the RF coherent detection, the output of the local oscillator in an optical coherent detection is not required to have the same phase as the incoming radiation. The principal advantages of a coherent receiver are: relative ease of amplification at an intermediate frequency and the fact that the signal-to-noise ratio can be significantly improved by simply raising the local oscillator power.
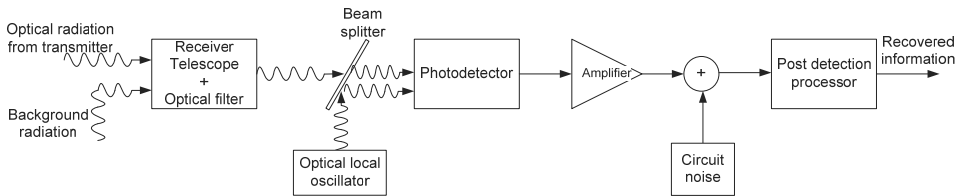
Fig. 5. The block diagram of a coherent detection optical receiver.

3.3 The atmospheric channel

An optical communications channel differs from the conventional Gaussian-noise channel, in that the channel input signal $x(t)$ represents power rather than amplitude. This leads to two constraints on the transmitted signal: i) $x(t)$ must be non-negative, and ii) the average value of $x(t)$ must not exceed a specified value $P_{max} \geq \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t) dt$. In contrast to the conventional channels, where the signal-to-noise ratio (SNR) is proportional to the power, in optical systems the received electrical power and the variance of the shot noise are proportional to A_d^2 and A_d, respectively; where A_d is the receiver detector area. Thus, for a shot noise limited optical system, the SNR is proportional to A_d. This implies that for a given transmit power; a higher SNR can be attained by using a large area detector. However, as A_d increases so does its capacitance, which has a limiting effect on the receiver bandwidth. The atmospheric channel consists of gases (see Table 3), and aerosols – tiny particles suspended in the atmosphere. Also present in the atmosphere are rain, haze, fog and other forms of precipitation. The amount of precipitation present in the atmosphere depends on the location (longitude and latitude) and the season. The highest concentration of particles is obviously near the Earth surface within the troposphere; this decreases with increasing altitude up through to the ionosphere (Gagliardi and Karp, 1995).

Constituent	Volume Ratio (%)	Parts Per Million (ppm)
Nitrogen (N_2)	78.09	
Oxygen (O_2)	20.95	
Argon (Ar)	0.93	
Carbon dioxide (CO_2)	0.03	
Water vapour (H_2O)		40-40,000
Neon (Ne)		20
Helium (He)		5.2
Methane (CH_4)		1.5
Krypton (Kr)		1.1
Hydrogen (H_2)		1
Nitrous oxide (N_2O)		0.6
Carbon monoxide (CO)		0.2
Ozone (O_3)		0.05
Xenon (Xe)		0.09

Table 3. The gas constituents of the atmosphere (AFGL, 1986).

Another feature of interest is the atmospheric turbulence. When radiation strikes the Earth from the Sun, some of the radiation is absorbed by the Earth's surface thereby heating up its (Earth's) surface air mass. The resulting mass of warm and lighter air then rises up to mix turbulently with the surrounding cooler air mass to create atmospheric turbulence. This culminates in small (in the range of 0.01 to 0.1 degrees) but spatially and temporally fluctuating atmospheric temperature (Killinger, 2002). With the size distribution of the atmospheric constituents ranging from sub-micrometres to centimetres, an optical field that traverses the atmosphere is scattered and or absorbed resulting in the following:

3.3.1 Power loss
For an optical radiation traversing the atmosphere, some of the photons are extinguished (absorbed) by the molecular constituents (water vapour, CO_2, fog, ozone etc) and their energy converted into heat while others experience no loss of energy but their initial direction of propagation are changed (scattering). The Beer-Lambert law describes the transmittance of an optical field through the atmosphere. The beam also spreads out while traversing the channel causing the size of the received beam to be greater than the receiver aperture. These factors, combined with others herein discussed are responsible for the difference between the transmitted and the received optical powers.

3.3.1.1 Atmospheric channel loss
The atmospheric channel attenuates the field traversing it as a result of absorption and scattering processes. The concentrations of matter in the atmosphere, which result in the signal attenuation vary spatially and temporally, and will depend on the current local weather conditions. For a terrestrial FSO link transmitting optical signal through the atmosphere, the received irradiance at a distance, L from the transmitter is related to the transmitted irradiance by the Beer-Lambert's law given as (Gagliardi and Karp, 1995):

$$\tau(\lambda, L) = \frac{P_R}{P_T} = \exp[-\gamma(\lambda)L], \tag{1}$$

where $\gamma(\lambda)$ and $\tau(\lambda, L)$ represent the total attenuation/extinction coefficient (m^{-1}) and the transmittance of the atmosphere at wavelength λ, respectively. The attenuation of the optical signal in the atmosphere is due to the presence of molecular constituents (gases) and aerosol. The attenuation coefficient is the sum of the absorption and the scattering coefficients from aerosols and molecular constituents of the atmosphere, it follows therefore that (Willebrand and Ghuman, 2002):

$$\gamma(\lambda) = \alpha_m(\lambda) + \alpha_a(\lambda) + \beta_m(\lambda) + \beta_a(\lambda) \tag{2}$$

The first two terms represent the molecular and aerosol absorption coefficients, respectively while the last two terms are the molecular and aerosol scattering coefficients respectively.

a) *Absorption* – This takes place when there is an interaction between the propagating photons and molecules (present in the atmosphere) along its path. Some of the photons are extinguished and their energies converted into heat (Pratt, 1969). The absorption coefficient depends very much on the type of gas molecules and their concentration (Gagliardi and Karp, 1995). Absorption is wavelength dependent and therefore selective. This leads to the atmosphere having transparent zones - range of wavelengths with minimal absorptions - referred to as the transmission windows. However, the wavelengths used in FSO are basically chosen to coincide with the atmospheric transmission windows (Bloom et al., 2003), resulting in the attenuation coefficient being dominated by scattering. Hence, $\gamma(\lambda) \cong \beta_a(\lambda)$.

b) *Scattering* – Results in angular redistribution of the optical field with and without wavelength modification. The scattering effect depends on the radius, r of the particles (fog, aerosol) encountered during propagation. One way of describing this is to consider the size parameter $x_o = 2\pi r/\lambda$. If $x_o \ll 1$, the scattering process is classified as Rayleigh scattering (Bates, 1984); if $x_o \approx 1$ it is Mie scattering and for $x_o \gg 1$, the scattering process can then be explained using the diffraction theory (geometric optics). The scattering process for different scattering particles present in the atmosphere is summarised in Table 4.

Type	Radius(μm)	Size Parameter x_o	Scattering Process
Air Molecules	0.0001	0.00074	Rayleigh
Haze particle	0.01 – 1	0.074 – 7.4	Rayleigh – Mie
Fog droplet	1 – 20	7.4 – 147.8	Mie – Geometrical
Rain	100 – 10000	740 – 74000	Geometrical
Snow	1000 – 5000	7400 –37000	Geometrical
Hail	5000–50000	37000 – 370000	Geometrical

Table 4. Typical atmospheric scattering particles with their radii and scattering process at $\lambda = 850$ nm

The fog particle size compares very much with the wavelength band of interest in FSO (0.5 μm – 2 μm). Thereby making fog a major photon scatterer and it contributes the most optical power attenuation. The Mie scattering will be described based on empirical formulae expressed in terms of the visibility range V in km. The visibility range is the distance that a parallel luminous beam travels through in the atmosphere until its intensity drops to 2% of its original value (Willebrand and Ghuman, 2002). The visibility is measured with an instrument called the transmissiometer. A common empirical model for Mie scattering is given by:

$$\beta_a(\lambda) = \frac{3.91}{V}\left(\frac{\lambda}{550}\right)^{-\delta} \tag{3}$$

where δ is given as:

$$
\begin{array}{cc}
\textbf{Kim model} & \textbf{Kruse model} \\
\delta = \begin{cases} 1.6 & V > 50 \\ 1.3 & 6 < V < 50 \\ 0.16V + 0.34 & 1 < V < 6 \\ V - 0.5 & 0.5 < V < 1 \\ 0 & V < 0.5 \end{cases} &
\delta = \begin{cases} 1.6 & V > 50 \\ 1.3 & 6 < V < 50 \\ 0.585V^{1/3} & V < 6 \end{cases}
\end{array} \tag{4}
$$

Given in Table 5 are the visibility range values under different weather conditions.

Weather Condition	Visibility Range (m)
Thick fog	200
Moderate fog	500
Light fog	770 – 1000
Thin fog/heavy rain (25mm/hr)	1900 – 2000
Haze/medium rain (12.5mm/hr)	2800 – 40000
Clear/drizzle (0.25mm/hr)	18000 – 20000
Very clear	23000 – 50000

Table 5. Weather conditions and their visibility range values

Recently, Al Naboulsi (al Naboulsi and Sizun, 2004) in his work came up with a simple relationship for advection and radiation fog attenuation in the 690 – 1550 nm wavelength range in the visibility range 50 – 1000 m as:

$$\alpha_{Advection}(\lambda) = \frac{0.11478\lambda + 3.8367}{V} \tag{5a}$$

$$\alpha_{Radiation}(\lambda) = \frac{0.18126\lambda^2 + 0.13709\lambda + 3.7502}{V} \tag{5b}$$

where λ is the wavelength in nm and the visibility *V* is in metres. The power loss due to rain and snow are so low compared to that due to the Mie scattering. But they still have to be accounted for in the link margin during the link budget analysis. A typical rainfall of 2.5 cm/hour could result in an attenuation of ~6 dB/km (Kim and Korevaar, 2001) while a typical value for attenuation due to light snow to blizzard is 3 dB/km to 30 dB/km (Willebrand and Ghuman, 2002). In early 2008 in Prague, Czech Republic, the fog attenuation was measured and compared with the empirical fog attenuation models. This result is shown in Fig. 6; with a visibility of less than 200 m – thick fog – the recorded fog attenuation is ~200 dB/km. All the empirical models provide a reasonable fit to the measured values with a maximum of about ±5 dB/km difference between any two empirical models.
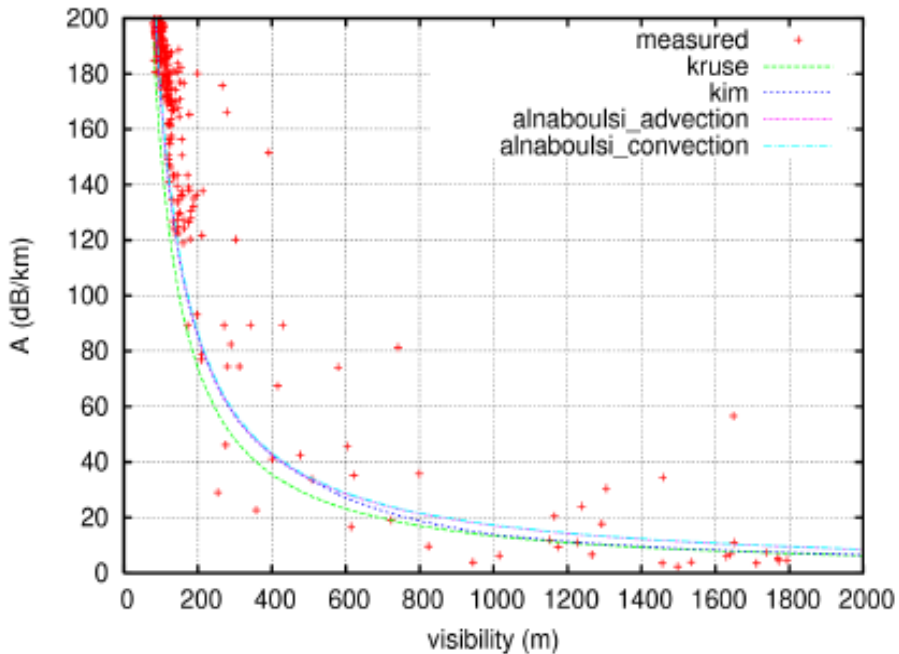
Fig. 6. Attenuation coefficient as a function of visibility range at λ = 830 nm (Grabner and Kvicera, 2009).

3.3.1.2 Beam divergence

One of the main advantages of FSO systems is the ability to transmit a very narrow optical beam, thus offering enhanced security. But due to diffraction, the beam spreads out. This results in a situation in which the receiver aperture is only able to collect a fraction of the beam and hence beam divergence loss.
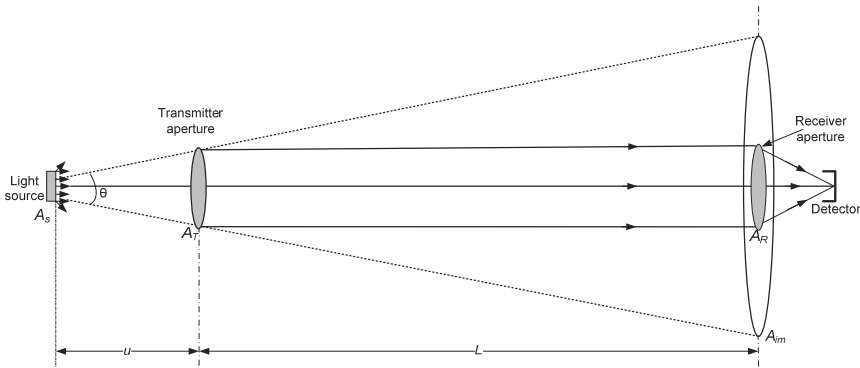
Fig. 7. Beam divergence

Considering the arrangement of a free-space optical communication link of Fig. 7, and by invoking the thin lens approximation to the diffuse optical source whose irradiance is represented by I_s, the amount of optical power focused on the detector is derived as (Gowar, 1993):

$$P_R = \frac{I_s A_T A_R}{L^2 A_s} \tag{6}$$

A_T and A_R are the transmitter and receiver aperture areas while A_s is the area of the optical source. This clearly shows that a source with high radiance I_s/A_s and wide apertures are required in order to increase the received optical power.

For a non-diffuse, small source such as the laser, the size of the image formed at the receiver plane is no longer given by the thin lens approximation; it is determined by diffraction at the transmitter aperture. The diffraction pattern produced by a uniformly illuminated circular aperture of diameter, d_T is known to consists of a set of concentric rings. The image size is said to be diffraction limited when the radius of the first intensity minimum or dark ring of the diffraction pattern becomes comparable in size with the diameter, d_{im} of the normally focussed image (Gowar, 1993). That is:

$$d_{im} = \frac{L}{u} d_s < 1.22 \frac{\lambda L}{d_T} \tag{7}$$

Therefore,

$$d_s < 1.22 \frac{\lambda u}{d_T} \approx 1.22 \frac{\lambda f}{d_T} \tag{8}$$

This equation shows that for diffraction to be the sole cause of beam divergence (diffraction limited), the source diameter, $d_s < 1.22 \frac{\lambda f}{d_T}$. Laser being inherently collimated and coherent normally produces a diffraction-limited image. The diffraction limited beam divergence angle in radian is given by $\theta_b \cong \frac{\lambda}{d_T}$.

If the transmitter and receiver effective antenna gains are respectively given by:

$$G_T = \frac{4\pi}{\Omega_b} \tag{9a}$$

$$G_R = \frac{4\pi A_T}{\lambda^2} \tag{9b}$$

And the free-space path loss is given by:

$$\mathcal{L} = \left(\frac{\lambda}{4\pi L}\right)^2 \tag{10}$$

Hence the received optical power is:

$$P_R = P_T \mathcal{L} G_T G_R \tag{11}$$

$$P_R = P_T \frac{4 A_R}{L^2 \Omega_b} \tag{12a}$$

$$P_R \cong P_T \left(\frac{4}{\pi}\right)^2 \frac{A_T A_R}{L^2 \lambda^2} \tag{12b}$$

where the radiation solid angle $\Omega_b \cong \frac{\pi \theta_b^2}{4}$. The diffraction limited beam spreading /geometric loss in dB is thus:

$$L_{Geom} = -10 \left[\log\left(\frac{A_T A_R}{L^2 \lambda^2}\right) + 2\log\left(\frac{4}{\pi}\right) \right] \tag{13}$$

The result given by (13) can be obtained by substituting $A_{im} = \theta_b L$ for the image size in $P_R = P_T \frac{A_R}{A_{im}}$. A beam expander of the type shown in Fig. 8, in which the diffracting aperture has been increased, could then be used to reduce the diffraction-limited beam divergence. Thereby reducing the beam divergence loss and increasing the received power in the process.
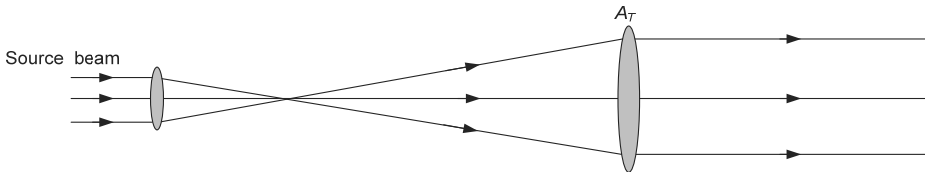

Fig. 8. Beam expander diagram

However for most practical sources, the beam divergence angle is usually greater than that dictated by diffraction. For a source with an angle of divergence θ, the beam size at a

distance L away is $(d_T + \theta L)$. The fraction of the received power to the transmitted power is therefore be given as:

$$\frac{P_R}{P_T} = \frac{A_R}{A_{Im}} = \frac{d_R^2}{(d_T + \theta L)^2}$$

(14)

And the geometric loss in dB thus becomes:

$$L_{Geom} = -20 \log\left[\frac{d_R}{(d_T + \theta L)}\right]$$

(15)

The beam spreading loss for the diffraction limited source given by (13) is expectedly lower than for the non-diffraction limited case given by (15), since the image size is smaller by d_T in the diffraction limited case.

From the foregoing, a source with very narrow divergence beam angle is preferable. It should however be mentioned that wide divergence angles are desirable in short range FSO links to ease the alignment requirement, compensate for building sway and eliminate the need for active tracking systems at the expense of increased geometric loss apparently. A typical FSO transceiver has optical beam divergence in the range of 2–10 mrad and 0.05–1.0 mrad (equivalent to a beam spread of 2–10 m, and 5 cm to 1 m, respectively at 1 km link range) for systems without and with tracking, respectively.

3.3.1.3 Optical and window loss

This type of loss includes losses due to imperfect lenses and other optical elements used in the design of both the transmitter and receiver. It accounts for the reflection, absorption and scattering due to the lenses in the system (Willebrand and Ghuman, 2002). The value of the optical loss L_O can be obtained from the component manufacturer. It apparently depends on the characteristics of the equipments and the quality of the lenses used. For FSO transceivers installed behind windows within a building, there exists an additional optical power loss due the window glass attenuation. Although (glass) windows allow optical signals to pass through them, they contribute to the overall power loss of the signal. Uncoated glass windows usually attenuate 4% per surface, because of reflection. Coated windows display much higher losses and its magnitude is wavelength dependent.

3.3.1.4 Pointing loss

Additional power penalty is usually incurred due to lack of perfect alignment of the transmitter and receiver. The resulting power loss is catered for by including pointing/misalignment loss, L_P in the link budget analysis. For short FSO links (< 1 km), this might not be an issue but for longer link ranges, this can certainly not be neglected. Misalignments could result from building sway or strong wind effect on the FOS link head stands.

3.3.2 The link budget

Based on the losses mentioned above, the received optical power in dBm can thus be obtained from the link budget equation as:

$$P_R(\lambda, L) = P_T(\lambda, 0) - 4.343L\beta_a(\lambda) - L_{Geom} - L_O - L_P - L_M \qquad (16)$$

The link margin, L_M is included in the link budget equation above to cater for other losses such as changes in specification when a faulty component is replaced, ageing of laser sources, attenuation due to rain, snow and so on. Figure 9 depicts the link range against available link margin at different values of visibility for a typical commercial FSO link whose parameters are tabulated in Table 6. In this figure, the Kim model is used in estimating the attenuation coefficient. By operating the link under consideration at a 5 dB link margin in clear atmosphere with over 30 km visibility, two data nodes at about 3 km apart and running at 155 Mbps can be reliably connected with an FSO system whose parameters are shown in Table 6.

Parameter	Typical Value
Receiver aperture diameter (d_R)	8 cm
Transmitter aperture diameter (d_T)	2.5 cm
Beam divergence (θ)	2 mrad
Modulation technique/Bit rate	On-OFF keying/155Mbps
Transmit power	14 dBm
Receiver sensitivity	-30 dBm
Optical loss (L_O)	1 dB
Pointing loss (L_P)	1 dB
Wavelength (λ)	850 nm

Table 6. Typical link budget parameters

One major importance of the link budget equation is in determining the achievable link range, for a given receiver sensitivity. The receiver sensitivity by the way represents the minimum amount of optical power needed for the system to achieve a specified level performance; for example a bit error rate of 10^{-9}. The receiver sensitivity depends on the modulation technique in use, the noise level, fading/scintillation strength and the data rate. Higher data rate simply implies shorter optical pulse duration, hence fewer photons that can be detected. The noise could be from a combination of background radiation, the detection process/quantum shot noise and the thermal noise caused by the thermal agitation of electrons in the receiver electronic components. The theoretical receiver sensitivity at any desired level of performance can be obtained from the analysis of Section 5.
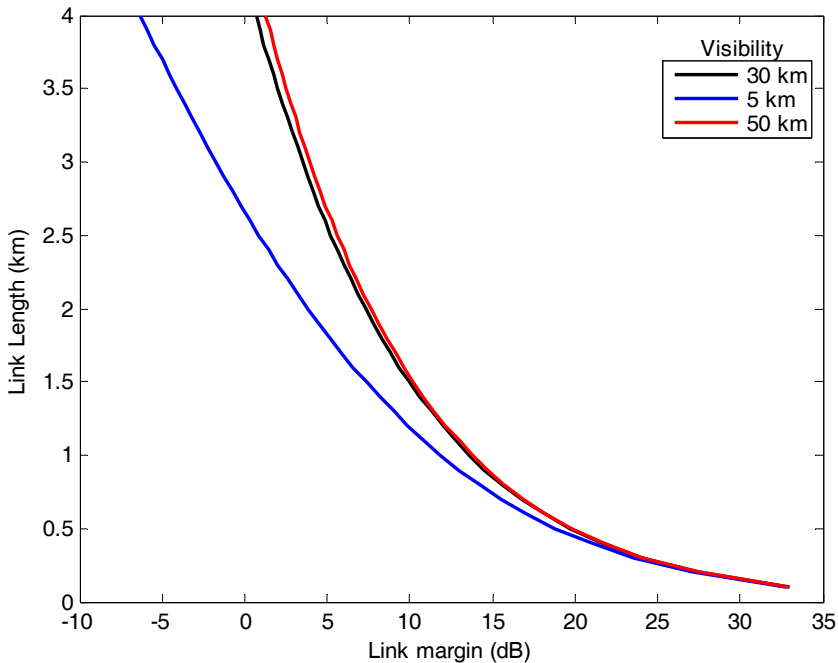
Fig. 9. Link length against available link margin for different visibility values.

3.3.3 The atmospheric turbulence effects

The temperature inhomogeneity of the atmosphere causes corresponding changes in the index of refraction of the atmosphere resulting in eddies, cells or air packets having varying sizes from ~0.1 cm to ~10 m. These air packets act like refractive prisms of varying indices of refraction. The propagating optical radiation is therefore fully or partially deviated depending on the relative size of the beam and the degree of temperature inhomogeneity along its path. Consequently, the optical radiation traversing the turbulence atmosphere experiences random variation/fading in its irradiance (scintillation) and phase. Familiar effects of turbulence include the twinkling of stars caused by random fluctuations of stars' irradiance and the shimmer of the horizon on a hot day caused by random changes in the optical phase of the light beam resulting in the reduced image resolution (Killinger, 2002). Atmospheric turbulence depends on i) atmospheric pressure/altitude, ii) wind speed, and iii) variation of index of refraction due to temperature inhomogeneity. Known effects of atmospheric turbulence include (Pratt, 1969):

a)   *Beam steering -* Angular deviation of the beam from its original LOS causing the beam to miss the receiver.
b)   *Image dancing -* The received beam focus moves in the image plane due to variations in the beam's angle of arrival.
c)   *Beam spreading -* Increased beam divergence due to scattering. This leads to a reduction in received power density.

d) *Beam scintillation -* Variations in the spatial power density at the receiver plane caused by small scale destructive interference within the optical beam.

e) *Spatial coherence degradation -* Turbulence also induces losses in phase coherence across the beam phase fronts. This is particularly deleterious for photomixing (e.g. in coherent receiver).

f) *Polarisation fluctuation* - This results from changes in the state of polarisation of the received optical field after passing through a turbulent medium. However, the amount of polarisation fluctuation is negligible for a horizontally travelling optical radiation in atmospheric turbulence (Karp et al., 1988).

3.3.3.1 Atmospheric turbulence model

Atmospheric turbulence results from random fluctuation of the atmospheric refractive index *n* along the path of a wave traversing the atmosphere. This refractive index fluctuation is the direct product of random variations in atmospheric temperature along the wave path. The random temperature changes themselves are a function of the altitude, *h* and the wind speed, *v*. Scintillation causes impairment and performance degradation for long range (> 1 km) atmospheric optical communication systems. The relationship between the temperature of the atmosphere and its refractive index is given by (Karp et al., 1988):

$$n = 1 + 77.6(1 + 7.52 \times 10^{-3}\lambda^{-2})\frac{P}{T_e} \times 10^{-6} \qquad (17)$$

where *P* is the atmospheric pressure in millibars, and T_e is the temperature in Kelvin.

The turbulence atmosphere can be described as containing loosely packed eddies/prisms of varying sizes and refractive indices. The smallest eddy size l_o is called the turbulence inner scale, with a value of a few millimetres, while the outer scale of turbulence L_o has its value running to several meters. According to the Taylor's 'frozen-in' model, the temporal variation in statistical properties of the turbulent atmosphere is caused by the airmass movement. Also, the turbulent eddies are fixed and only vary with the wind moving perpendicularly to the direction of the traversing wave. The temporal coherence time τ_o of atmospheric turbulence is known to be in the order of millisecond. This value is very large compared to typical data symbol duration. Hence the terrestrial FSO channel suffers from slow fading.

Since only the intensity modulation, direct detection laser communication systems are discussed here, the turbulence effect of concern is the intensity fluctuation of the laser beam traversing the atmosphere. The strength of the irradiance fluctuation in a turbulent medium is given by the variance of the log intensity, *l* (also called the Roytov parameter σ_l^2) and the transverse coherence length of a field travelling through a turbulent channel is denoted by ρ_o. Over the range $l_o \leq \sqrt{\lambda L} \leq L_o$ these parameters are defined as (Osche, 2002):

$$\sigma_l^2 = 2.25(2\pi/\lambda)^{7/6} \int_0^\infty C_n^2(x)\,(x/L)^{5/6}(L-x)^{5/6}dx \qquad (18)$$

$$\rho_o \approx \sqrt{\lambda L} \qquad (19)$$

where $C_n{}^2$ is the refractive index structure constant (which characterizes the strength of refractive index variation in the medium). A commonly used model for $C_n{}^2$ is the Hufnagel-Valley (H-V) model described by the following (Andrews et al., 2001):

$$C_n^2(h) = 0.00594(v/27)^2(10^{-5}h)^{10} \exp(-h/1000) + 2.7 \times 10^{-16} \exp(-h/1500) + \hat{A}\exp(-h/100) \tag{20}$$

$\hat{A}$ is taken as the nominal value of $C_n{}^2(0)$ at ground level in $m^{-2/3}$. Generally, the structure parameter is assumed constant for a horizontal link and ranges from 10^{-15} $m^{-2/3}$ for weak to 10^{-12} $m^{-2/3}$ for strong turbulence regimes.

Considering single scattering characterized weak turbulence and assuming the log intensity l of laser light traversing the turbulent atmosphere to be normally distributed, that is $l \sim N(-\sigma_l^2/2, \sigma_l^2)$, then the probability density function (pdf) of the laser beam intensity, $I = I_o\exp(l)$, is given by:

$$p(I) = \frac{1}{\sqrt{2\pi\sigma_l^2}} \frac{1}{I} \exp\left\{-\frac{[\ln(I/I_o) + \sigma_l^2/2]^2}{2\sigma_l^2}\right\} \tag{21}$$

where I_o is the mean received intensity without turbulence. The normalised variance of the intensity σ_N^2 is derived as follows:

$$\sigma_N^2 = E[I^2] - (E[I])^2/(E[I])^2 = \exp(\sigma_l^2) - 1 \tag{22}$$

The turbulence model discussed thus far is the lognormal turbulence, it is only valid for the weak turbulence with $\sigma_N{}^2 < 1.2$. For $\sigma_N^2 \geq 1.2$, saturation sets in and the model no longer holds. Turbulence induced irradiance fluctuation can enter saturation due to one or a combination of increased C_n, link length and reduced wavelength. Also, when multiple scatterings are experienced especially in longer link ranges, the incident wave becomes increasingly incoherent and log normal model becomes invalid. Though not discussed here, another model which has a wider range of validity but lacks the mathematical simplicity of lognormal is the gamma-gamma turbulence model. Moreover, in the limit of strong irradiance fluctuations (i.e. in saturation regime and beyond) where the link length spans several kilometres, the number of independent scatterings becomes large (Karp et al., 1988). In the saturation regime, irradiance fluctuation is believed to follow the negative exponential distribution.

4. Noise Sources

Background noise: This is due to radiations from both the sky (extended source) and the Sun (localised source). Background radiation from other celestial bodies such as stars and reflected background radiation are assumed too weak to be considered for terrestrial FSO links, they however contribute significantly to background noise in deep space FSO systems.

The irradiance (power per unit area) expressions for both the extended and the localised background sources are given by the following equations:

$$I_{sky} = N(\lambda)\Delta\lambda\pi\Omega^2/4 \tag{23}$$

$$I_{sun} = W(\lambda)\Delta\lambda \tag{24}$$

where $N(\lambda)$ and $W(\lambda)$ are the spectral radiance of the sky and spectral radiant emittance of the Sun, respectively, $\Delta\lambda$ is the bandwidth of the optical BPF at the receiver, and Ω is the receiver FOV in radian. By carefully choosing a receiver with a very narrow FOV and $\Delta\lambda$, the impact of background noise can be greatly reduced. Optical BPF in the form of coatings on the receiver optics/telescope with $\Delta\lambda < 1$ nm are now readily available. Empirical values of $N(\lambda)$ and $W(\lambda)$ under different observation conditions are also available in reference (Gagliardi and Karp, 1995). The background noise is a shot noise and its variance is given by:

$$\sigma_{Bg}^2 = 2qB\Re\left(I_{sky} + I_{sun}\right) \tag{25}$$

where B is the system electrical bandwidth and $\Re = \eta q\lambda/hc$ is the photodetector responsivity, η is the detector quantum efficiency, q is the electronic charge; h and c represent the Plank's constant and the speed of light in vacuum, respectively.

Quantum noise: A shot noise due to the statistical nature of the optical detection process. Its value is usually very small with variance:

$$\sigma_{Qtm}^2 = 2qB\Re I \tag{26}$$

Thermal noise: This is the noise caused by the thermal fluctuations of electrons in a receiver circuit of equivalent resistance R_L, and temperature T_e. Its variance is given by:

$$\sigma_{Th}^2 = \frac{4kT_eB}{R_L} \tag{27}$$

The dark current and the relative intensity noise are usually so small and negligible. The total noise variance is thus given as:

$$\sigma^2 = \sigma_{Qtm}^2 + \sigma_{Bg}^2 + \sigma_{Th}^2 \tag{28}$$

The major challenges associated with the optical wireless communication systems are summarised in Table 7.

Challenge	Causes	Effects	Mitigation Approach	Indoor / FSO
Inter symbol interference (ISI)	Multipath Propagation	Poor transmission quality (high BER)	Channel equalization	Indoor
			Forward error control (FEC)	
		Multipath distortion or dispersion	Spread spectrum techniques	
			Multiple subcarrier Modulation (*More bandwidth efficient than a single-carrier system*)	
		Reduced date rates	OFDM, MSM	
			Multi-beam transmitter	
			FOV controlling	
Safety	Laser Radiation	Damage to eyes and skin	Power efficient modulation schemes: - PPM, DPIM, etc.	Both
			Use LED, Class 1 lasers, and 1550 nm wavelength	
Noise	Dark current noise	Low Signal-to-noise ratio and high BER	Optical and electrical filtering	Both
	Shot noise			
	Background noise		Pre-amplification	
	Thermal noise		FEC	
	Relative intensity noise		Low noise post detection amplifier	
	Excess noise (with APD)		Small FOV lasers	
	ASE (only if optical amplifier is used)		Optical filter	
Turbulence	Random refractive index variation	Phase and intensity fluctuations (scintillation)	FEC (LDPC, Turbo codes) Robust modulation: SIM, PPM MIMO Diversity reception (temporal and spatial)	FSO
		Image dancing		
		Spatial coherence degradation		

		Beam spreading	Adaptive optics	
Reflection index	Different materials	Higher losses due to reflection on surfaces	Higher transmit power	Indoor
Blocking	Furniture	Temporary link outage	Diffuse link	Both
	Moving objects		Cellular system	
	Walls		Multi-beam	
	Birds		Hybrid FSO/RF	
Weather effects	Fog, Rain, Gases, Smoke, Aerosols	Attenuations, Scattering Link outage	Higher transmit power Hybrid FSO/RF	FSO
Pointing, Acquisition and Tracking (PAT)	Mobile link heads Building sway	Temporary/perma nent link outage Power loss	Hybrid FSO/RF Active tracking Adaptive optics (beam steering and tracking)	FSO

Table 7. Challenges in Optical Wireless Communications

5. Modulation Techniques

There exist different types of modulation schemes that are suitable for optical wireless communication systems, one of such is the family of pulse modulation techniques shown in Fig. 10. Since the average emitted optical power is limited, the different modulation techniques are usually compared in terms of the average received optical power required to achieve a desired bit error rate at a given data rate. A power efficient modulation scheme is desirable in order to maximise the ratio of peak to average power. Here, the performance analysis of an FSO system based on the following modulation techniques: On-off keying (OOK), pulse position modulation (PPM) and subcarrier intensity modulation (SIM) for non-ideal channels will be highlighted.
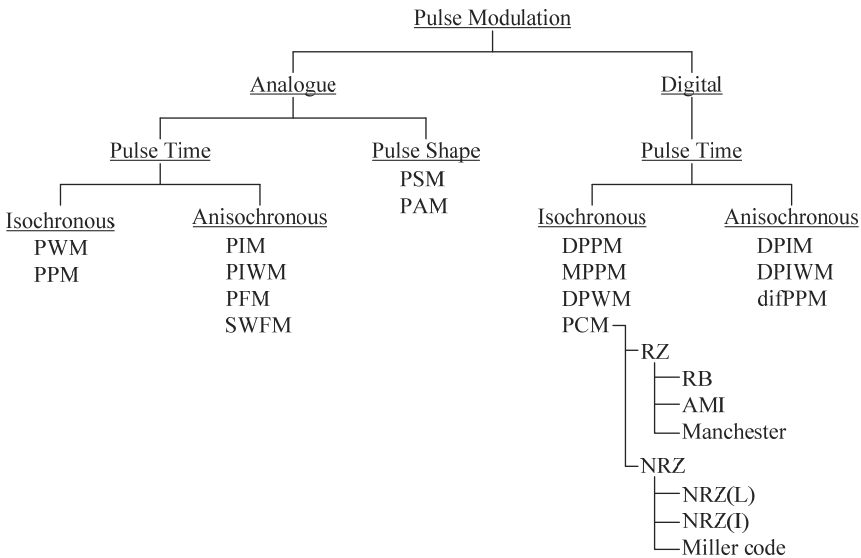
Pulse Modulation

Analogue

Digital

Pulse Time

Pulse Shape
PSM
PAM

Pulse Time

Isochronous
PWM
PPM

Anisochronous
PIM
PIWM
PFM
SWFM

Isochronous
DPPM
MPPM
DPWM
PCM

Anisochronous
DPIM
DPIWM
difPPM

RZ
  RB
  AMI
  Manchester

NRZ
  NRZ(L)
  NRZ(I)
  Miller code

Fig. 10. Pulse modulation tree.

5.1 On-Off Keying

The OOK signalling is the dominant modulation scheme employed in terrestrial wireless optical communication systems. This is primarily due to its simplicity and resilience to laser nonlinearity. OOK can use either non-return-to-zero (NRZ) or return-to-zero (RZ) pulse formats. In NRZ-OOK, an optical pulse of peak power $\alpha_e P_T$ represents a digital symbol '0' while the transmission of an optical pulse of peak power P_T represents a digital symbol '1'. The optical source extinction ratio α_e lies in the range $0 \leq \alpha_e < 1$. The finite duration of the optical pulse is the same as the symbol duration T. With OOK-RZ, the pulse duration is lower than the bit duration, giving an improvement in power efficiency over NRZ-OOK at the expense of an increased bandwidth requirement. Without any loss of generality, the receiver area can be normalised to unity such that the optical power can henceforth be represented by the optical intensity I. If $\Re$ represents the responsivity of the PIN photodetector, the received signal in an OOK modulated FSO system becomes:

$$i(t) = \Re I \left[1 + \sum_{j=-\infty}^{\infty} d_j g(t - jT) \right] + n(t) \tag{29}$$

where $n(t) \sim N(0, \sigma^2)$ is the additive white Gaussian noise and d_j = [-1,0]. In all the analyses that follow, the extinction ratio is assumed equal to zero unless otherwise stated. At the receiver, the received signal is fed into a threshold detector which compares the received signal with a pre-determined threshold level. A digital symbol '1' is assumed received if the received signal is above the threshold level and '0' otherwise. The probability of error is therefore given as:

$$P_e = p(0) \int\limits_{i_{th}}^{\infty} p(i/0)di + p(1) \int\limits_{0}^{i_{th}} p(i/1)di \qquad (30)$$

where the marginal probabilities are defined as:

$$p(i/0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-i^2/2\sigma^2) \qquad (31)$$

$$p(i/1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(i-\Re I)^2}{2\sigma^2}\right] \qquad (32)$$

For equiprobable symbols, $p(0) = p(1) = 0.5$, the optimum threshold point is at $i_{th} = 0.5\Re I$. And the conditional probability of error reduces to:

$$P_{ec} = Q\left(\frac{i_{th}}{\sigma}\right) \qquad (33)$$

where $Q(x) = 0.5\mathrm{erfc}(x/\sqrt{2})$. But in the presence of atmospheric turbulence, the threshold level is no longer fixed midway between the signal levels representing symbols '1' and '0'. The marginal probability $p(i/1)$ is then modified by averaging equation (32) over the scintillation statistics to arrive at equation (34). Note that scintillation does not occur when no pulse is transmitted.

$$p(i/1) = \int\limits_{0}^{\infty} p(i/1, I)p(I)dI \qquad (34)$$

Assuming equiprobable symbol transmission and invoking the maximum a posteriori symbol-by-symbol detection, the likelihood function becomes (Popoola et al., 2008):

$$\Lambda = \int\limits_{0}^{\infty} \exp\left[\frac{-(i-\Re I)^2 - i^2}{2\sigma^2}\right] p(I)dI \qquad (35)$$

The threshold level i_{th} is obtained from (35) with $\Lambda = 1$. Based on the log normal turbulence model, the plot of i_{th} for different levels of turbulence is shown in Fig. 11.
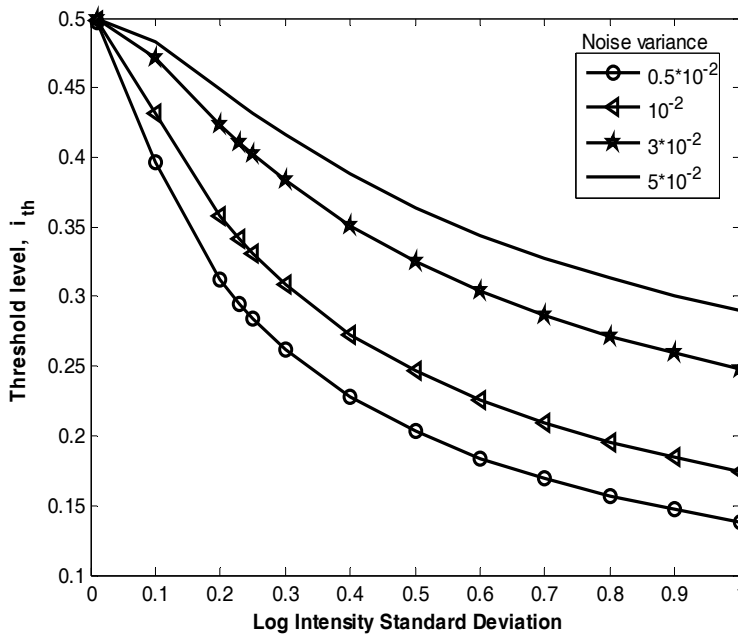
Fig. 11. OOK threshold level against the log intensity standard deviation for a range of turbulence levels.

The threshold level is observed to approach the 0.5 value as the scintillation level approaches zero. As an illustration, at a turbulence level $\sigma_l^2 = 0.2$, the probability of bit error P_e, obtained from the combination of (30), (31) and (34) is plotted against the normalised SNR = $(\Re E[I])^2/\sigma^2$ in Fig. 12, the value of i_{th} used for the adaptive threshold level graph is obtained from the solution of equation (35). From this figure, the effect of using a fixed threshold level in fading channels results in a BER floor. The values of which depend on the fixed threshold level and turbulence induced fading strength. With an adaptive threshold, there is no such BER floor and any desired level of BER can thus be realised.
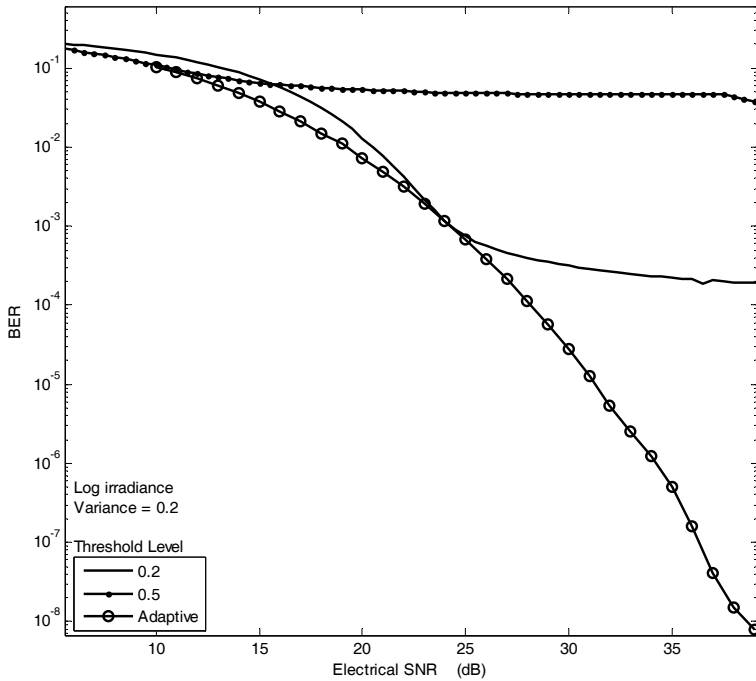
Fig. 12. BER performance of OOK based FSO in atmospheric turbulence, $\sigma_l^2 = 0.2$.

In Fig. 13, the BER is again plotted against the normalised SNR at various levels of scintillation including when the threshold is fixed at 0.5. This is intended to show the effect of turbulence strength on the amount of SNR required to maintain a given error performance level. With a fixed threshold, the BER reaches a floor at a BER that is greater than 10^{-4}, meaning that a lower BER is not achievable at the specified low scintillation level. From this graph, it can be inferred that atmospheric turbulence: i) causes SNR penalty, for example ~26 dB of SNR is needed to achieve a BER of 10^{-6} due to a very weak scintillation of strength $\sigma_l^2 = 0.25^2$, this however increases by over 20 dB as the scintillation strength increases to $\sigma_l^2 = 0.7^2$; and ii) implies that adaptive threshold will be required to avoid a BER floor in the system performance.
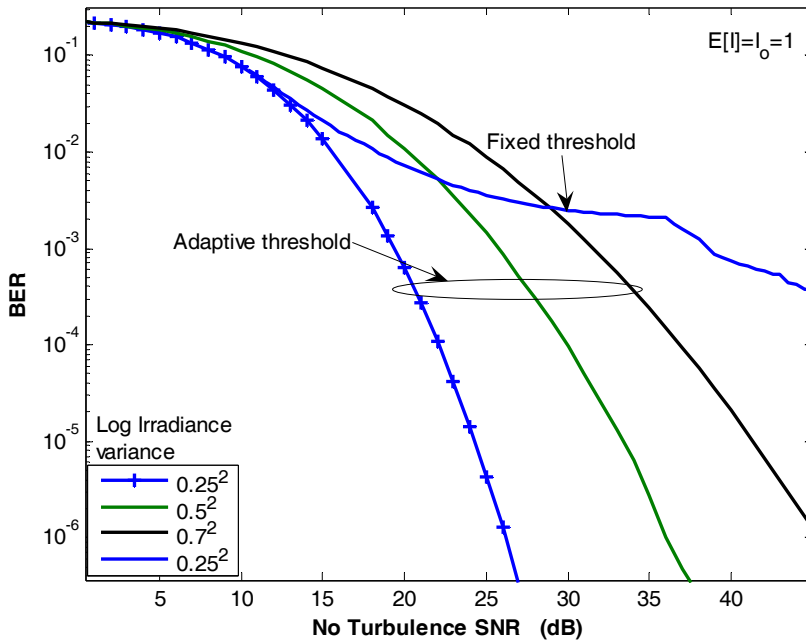
Fig. 13. BER against the SNR for fixed and adaptive thresholds and various levels of scintillation.

The results of this section illustrate that for the OOK modulated FSO system to perform at its best, the receiver will require knowledge of both the fading strength and the noise level. This can be resolved by integrating into the system an intensity estimation network which can predict the scintillation level based on past events. The implementation of this is not trivial, and as such, commercial FSO designers tend to adopt the fixed threshold approach and include a sufficiently large link margin in the link budget to cater for turbulence induced fading.

5.2 Pulse Position Modulation

This is an orthogonal modulation technique and a member of the pulse modulation family (see Fig. 10). The PPM modulation technique improves on the power efficiency of OOK but at the expense of an increased bandwidth requirement and greater complexity. In PPM, each block of $\log_2 M$ data bits is mapped to one of M possible symbols. Generally, the notation M-PPM is used to indicate the order. Each symbol consists of a pulse of constant power P_T occupying one slot, along with M-1 empty slots. The position of the pulse corresponds to the decimal value of the $\log_2 M$ data bits. Hence, the information is encoded by the position of the pulse within the symbol. The slot duration T_s is related to the bit duration by the following expression:

$$T_s = \frac{T \log_2 M}{M}$$
(36)

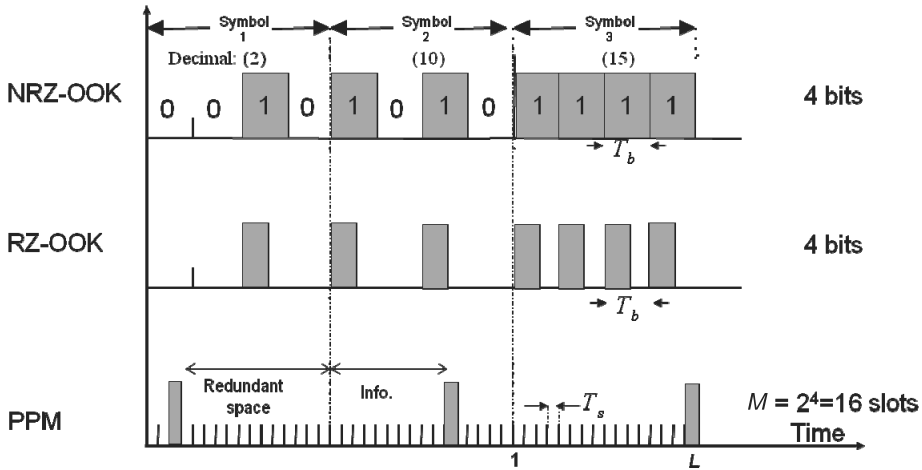The transmitted waveforms for 16-PPM and OOK are shown in Fig. 14.

Fig. 14. Transmitted waveforms for OOK and 16-PPM

A PPM receiver will require both slot and symbol synchronisation in order to demodulate the information encoded on the pulse position. Nevertheless, because of its superior power efficiency, PPM is an attractive modulation technique for optical wireless communication systems particularly in deep space laser communication applications. Assuming that a complete synchronisation is maintained between the transmitter and receiver at all times, the optical receiver detects the transmitted signal by attempting to determine the optical energy in each possible time slot. In direct photodetection this is equivalent to "counting" the number of released electrons in each T_s interval. The photo count per PPM slot is given by:

$$K_s = \frac{\eta \lambda P_R T_s}{hc}$$
(37)

where P_R is the received optical power during a slot duration. An APD could be used to afford an increase in the amount of photon count per PPM slot but unfortunately, the photo-multiplication process that governs the generation of the secondary electrons is a random process. This implies that a large photo-multiplication gain will eventually lead to a large noise factor and an error prone performance. For a moderately high received signal as is the case in commercial and short range FSO systems, the BER conditioned on K_s is given by (Gagliardi and Karp, 1995):

$$P_{ec} = Q\left(\sqrt{\frac{(\bar{g}q)^2 K_s^2}{(\bar{g}q)^2 F(K_s + 2K_{Bg}) + 2\sigma_{Th}^2}}\right) \tag{38}$$

where the parameters are defined as:

$K_{Bg} = \eta\lambda P_{Bg} T_s/hc$	Average photon count per PPM slot due to the background Radiation of power P_{Bg}
$\bar{g}$	Average APD gain
q	Electronic charge
$F \approx 2 + \zeta\bar{g}$	Noise factor of the APD
ζ	APD ionisation factor
$\sigma_{Th}^2 = (2\kappa T_e q/R_L)(T_s)$	Equivalent thermal noise count within a PPM slot duration.
$R_b = 1/T$	Bit rate.
κ	Boltzmann's constant
R_L	Equivalent load resistance

In the presence of log normal atmospheric turbulence, the unconditional BER for a binary PPM modulated FSO obtained by averaging (38) over the scintillation statistics can be approximated as (Kiasaleh, 2005):

$$P_e \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{n} w_i Q\left(\frac{\exp\left(2(\sqrt{2}\sigma_k x_i + m_k)\right)}{F\exp(\sqrt{2}\sigma_k x_i + m_k) + K_n}\right) \tag{39}$$

where $[w_i]_{i=1}^n$ and $[x_i]_{i=1}^n$ are the weight factors and the zeros of an nth order Hermite polynomial. $K_n = \left(2\sigma_{Th}^2/(\bar{g}q)^2\right) + 2FK_{Bg}$ and $\sigma_k^2 = \ln(\sigma_N^2 + 1)$. It is noteworthy that the fluctuation of the mean count K_s is brought about by the atmospheric turbulence and its ensemble average is given by:

$$E[K_s] = \exp\left(\frac{\sigma_k^2}{2} + m_k\right) \tag{40}$$

For an M-PPM system, the BER denoted by P_e^M has an upper bound given by (Kiasaleh, 2005):

$$P_e^M \leq \frac{M}{2\sqrt{\pi}} \sum_{i=1}^{n} w_i \mathrm{Q}\left(\frac{\exp\left(2\left(\sqrt{2}\sigma_k x_i + m_k\right)\right)}{F\exp\left(\sqrt{2}\sigma_k x_i + m_k\right) + K_n} \right) \tag{41}$$

The BER of a binary PPM modulated FSO is shown in Fig. 15 for different levels of scintillation.
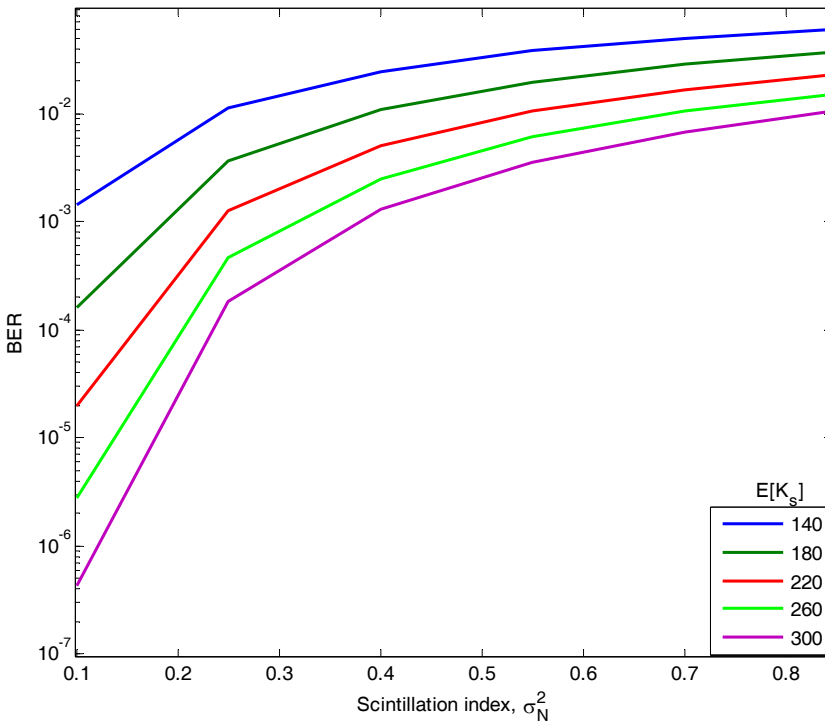
Fig. 15. Binary PPM BER as a function of scintillation index for K_{Bg} = 10; T_e = 300 K, ζ = 0.028, R_b = 155 Mbps and $\bar{g}$ = 150.

As expected, an increase in the atmospheric scintillation results in an increase in the required signal level to achieve a given BER. Increasing the signal strength can be used to minimise the scintillation effect at a low scintillation index, but as turbulence strength increases, it is observed that the BERs tend towards a high asymptotic value.

5.3 Subcarrier Intensity Modulation

(a)

(b)
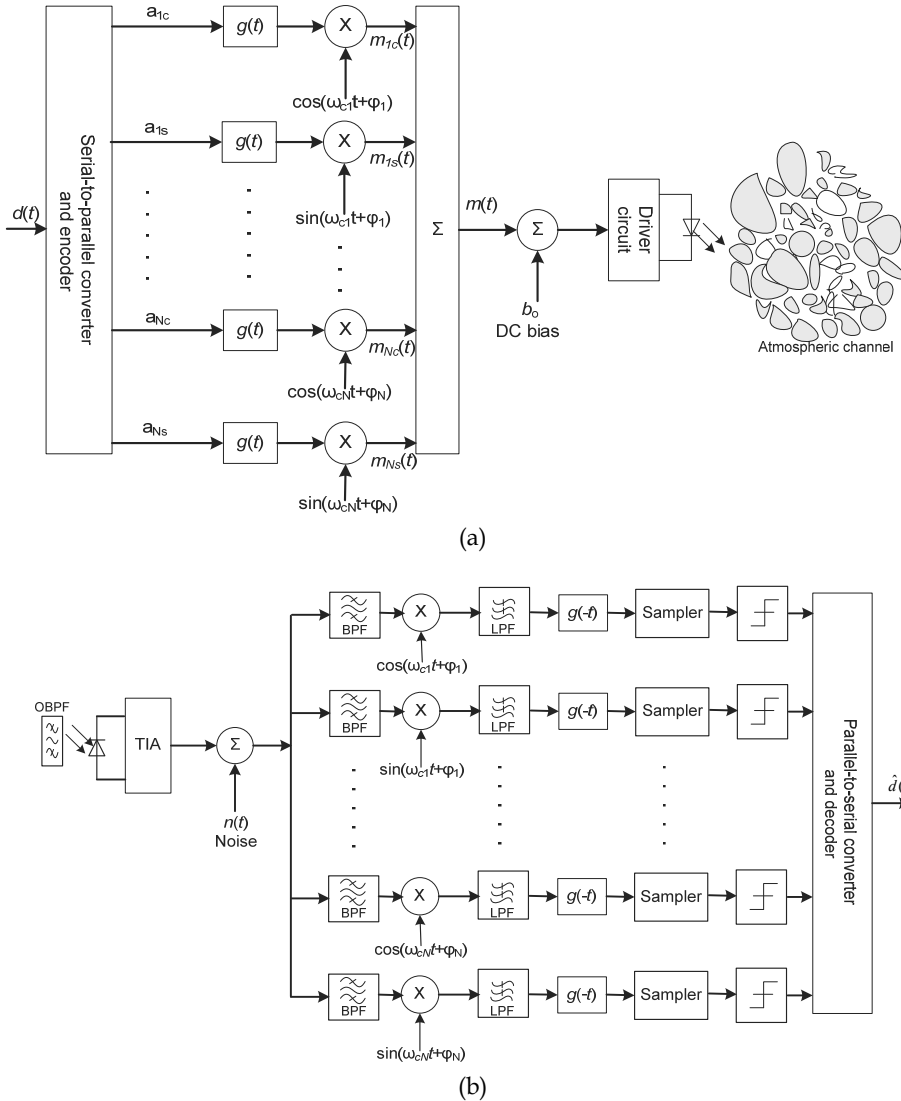
Fig. 16. Block diagram of SIM: (a) transmitter, and (b) receiver. TIA- trans-impedance

In optical SIM, an RF subcarrier signal pre-modulated with the source data $d(t)$ is used to modulate the intensity of the optical carrier - a continuous wave laser diode (Popoola and Ghassemlooy, 2009). Figure 16 illustrates the system block diagram of a SIM-FSO with N subcarriers. The serial-to-parallel converter distributes the incoming data across the N-subcarriers. Each subcarrier carries a reduced symbol rate but the aggregate must be equal to the symbol rate of $d(t)$. Another obvious possibility not shown in the figure is to have

different users occupying the N different subcarriers. Prior to modulating the laser irradiance, the source data $d(t)$ is modulated onto the RF subcarriers. For the M-PSK subcarrier modulation shown in Fig. 16, the encoder maps each subcarrier symbol onto the symbol amplitude $\{a_{ic}, a_{is}\}_{i=1}^{N}$ that corresponds to the constellation in use. Since the subcarrier signal $m(t)$ is sinusoidal having both positive and negative values, a DC level b_0 is added to $m(t)$ before it is used to directly drive the laser diode. This is to ensure that the bias current is always equal to or greater than the threshold current - a prerequisite for the avoidance of signal clipping.

The following gives the general expression for $m(t)$ in the N-SIM-FSO system:

$$m(t) = \sum_{i=1}^{N} m_i(t) \tag{42}$$

During symbol duration, each RF subcarrier signal is generally represented by:

$$m_i(t) = g(t)a_{ic}\cos(\omega_{ci}t + \varphi_i) + g(t)a_{is}\sin(\omega_{ci}t + \varphi_i) \tag{43}$$

where $g(t)$ is the pulse shaping function, the subcarrier angular frequency and phase are represented by $[\omega_{ci}, \varphi_i]_{i=1}^{N}$. It follows that each subcarrier can be modulated by any standard RF digital/analogue modulation techniques such as QAM, M-PSK, M-FSK and M-ASK. SIM benefits from advances made in RF modulators/demodulators, stable oscillators and narrow band filters.

Using a direct detection at the receiver, the incoming optical radiation is converted into an electrical signal. This is followed by a standard RF demodulator to recover the transmitted symbol as shown in Fig. 16(b). The received signal is modelled as:

$$i(t) = \Re I[1 + \xi m(t)] + n(t) \tag{44}$$

where ξ is the modulation depth/index. The electrical band pass filters (BPF) performs the following functions: selection of the individual subcarrier for demodulation, reduction of the noise power and suppression of any slow varying RI component present in the received signal. For a subcarrier at ω_{ci} the received signal is:

$$i(t) = I_{comp} + Q_{comp} \tag{45}$$

where

$$I_{comp} = \Re I\xi g(t)a_{ic}\cos(\omega_{ci}t + \varphi_i) + n_I(t) \tag{46a}$$

$$Q_{comp} = -\Re I\xi g(t)a_{is}\sin(\omega_{ci}t + \varphi_i) + n_Q(t) \tag{46b}$$

$n_I(t)$ and $n_Q(t)$ are the independent AWGN with zero mean and variance σ². By down converting the quadrature components I_{comp} and Q_{comp} by the reference signals $\cos\omega_c t$ and

$\sin\omega_c t$, respectively, the following conditional BER expressions (47a-d) are obtained (Proakis, 2004):

$$P_{ec} = Q\left(\sqrt{\gamma(I)}\right) \qquad \text{for BPSK} \tag{47a}$$

$$P_{ec} \approx \frac{2}{\log_2 M} Q\left(\sqrt{(\log_2 M)\gamma(I)}\,\sin(\pi/M)\right) \qquad \text{for } M\text{-PSK}, M \geq 4 \tag{47b}$$

$$P_{ec} = 0.5\exp(-0.5\gamma(I)) \qquad \text{For DPSK} \tag{47c}$$

$$P_{ec} = \frac{2(1 - 1/\sqrt{M})}{\log_2 M} Q\left(\sqrt{\frac{3\log_2 M\,\gamma(I)}{2(M-1)}}\right) \qquad \text{for } M\text{-QAM}, \log_2 M \text{ even} \tag{47d}$$

The SNR per bit and the average power are respectively given as:

$$\gamma(I) = \frac{(\xi\Re I)^2 P_m}{\sigma^2} \tag{48}$$

$$P_m = \frac{A^2}{2T}\int_0^T g^2(t)\,dt \tag{49}$$

To illustrate the system performance in atmospheric turbulence, the BPSK-SIM is considered. Since each subcarrier is demodulated separately and independently, the unconditional BER per subcarrier is obtained by averaging (47a) over the irradiance fluctuation statistics to obtain:

$$P_e = \int_0^\infty Q\left(\sqrt{\gamma(I)}\right) p(I)\,dI \tag{50}$$

$$P_e = \frac{1}{\sqrt{\pi}}\sum_{i=1}^n w_i Q\left[\sqrt{K_0}\exp\left(K_1\left(\sqrt{2}\sigma_l x_i - \sigma_l^2/2\right)\right)\right] \tag{51}$$

where $[w_i]_{i=1}^n$ and $[x_i]_{i=1}^n$ are the weight factors and the zeros of an nth order Hermite polynomial. The values of K_1 and K_0 are given in Table 8 for different noise limiting conditions:

	Performance Limiting conditions			
	Quantum limit	Thermal noise	Background noise	Thermal and background noise
K_0	$\dfrac{\xi^2 \Re I_o P_m}{2qR_b}$	$\dfrac{(\xi\Re I_o)^2 P_m R_L}{4kT_e R_b}$	$\dfrac{(\xi I_o)^2 \Re P_m}{2qR_b(I_{sky} + I_{sun})}$	$\dfrac{(\xi\Re I_o)^2 P_m}{(\sigma_{Bg}^2 + \sigma_{Th}^2)}$
K_1	0.5	1	1	1

Table 8. Values of K_1 and K_0 shown for different noise limiting conditions

To keep the continuous wave laser source within its dynamic range, the condition $|\xi m(t)| \leq 1$ must always hold. This condition places an upper bound on the amplitude of each subcarrier for a given value of ξ. It also implies that the use of multi-user-SIM (subcarrier multiplexing) incurs a $20\log N$ dB reduction in $\gamma(l)$. Consequently, multiple subcarriers should only be considered when the need for increased capacity outweighs the accompanying power penalty.

Based on the simulation parameters of Table 9, the BER defined in (51) is plotted against the receiver sensitivity in Fig. 17 with $\xi = N = 1$ under the different noise conditions. The figure illustrates clearly that for an FSO link with a suitable optical BPF and a narrow FOV detector, the system performance is limited by thermal noise. Moreover, under this thermal noise limited condition, the SIM-FSO still requires about 30 dB of SNR more, compared with the theoretical quantum limit.
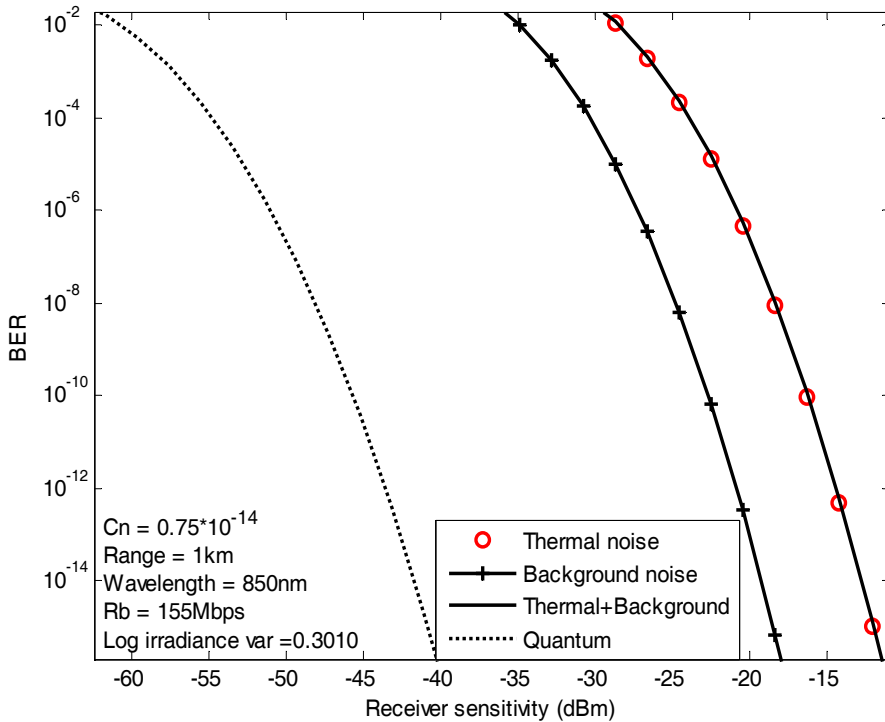

Fig. 17. The BER against the receiver sensitivity under different noise sources, with R_b = 155 Mpbs, σ_I^2 = 0.3 and ξ = 1.

Parameter	Value
Symbol rate R_b	155 Mbps
Spectral radiance of the sky $N(\lambda)$	10^{-3} W/cm²μmSr
Spectral radiant emittance of the sun $W(\lambda)$	0.055 W/cm²μm
Optical band-pass filter bandwidth $\Delta\lambda$ @ λ = 850 nm	1 nm
PIN photodetector field of view (FOV)	0.6 rad
Radiation wavelength λ	850 nm
Number of subcarriers N	1
Link range L	1 km
Index of refraction structure parameter C_n^2	0.75 m$^{-2/3}$
Load resistance R_L	50 Ω
PIN photodetector responsivity R	1
Operating temperature T_e	300 K

Table 9. Simulation parameters

6. Conclusion

In this chapter, the terrestrial FSO communication system has been discussed right from its basics to error performance based on OOK, PPM and SIM modulation schemes. The properties of the atmospheric channel have also been highlighted in terms of signal attenuation and scintillation. On the evidence of increasing deployment of FSO systems for civil applications, it is fair to say that the technology has now moved on from being researchers' pastime and bread board technology into a potent complementary access network technology. The provision of long range FSO links covering several kilometres with 99.999 % availability across all weather conditions however remains a daunting challenge.

7. References

AFGL (1986) AFGL Atmospheric Constituent Profiles (0–120 km), AFGL-TR-86-0110, Air Force Geophysics Laboratory, Hanscom Air Force Base, Massachusetts,.

AL NABOULSI, M. & SIZUN, H. (2004) Fog attenuation prediction for optical and infrared waves. *Optical Engineering,* 23**,** 319-329.

ANDREWS, L. C., PHILLIPS, R. L. & HOPEN, C. Y. (2001) *Laser beam scintillation with applications,* Bellingham, SPIE.

BATES, D. R. (1984) Rayleigh scattering by air. *Planetary Space Science,* 32**,** 785-790.

BLOOM, S., KOREVAAR, E., SCHUSTER, J. & WILLEBRAND, H. (2003) Understanding the performance of free-space optics. *Journal of Optical Networking,* 2**,** 178-200.

GAGLIARDI, R. M. & KARP, S. (1995) *Optical Communications,* New York, John Wiley.

GOODWIN, F. E. (1970) A review of operational laser communication systems. *Proceedings of IEEE,* 58**,** 1746-1752.

GOWAR, J. (1993) *Optical communication systems*, Prentice Hall.

GRABNER, M. & KVICERA, V. (2009) Experimental study of atmospheric visibility and optical wave attenuation for free-space optics communications. *http://ursi-france.institut-telecom.fr/pages/pages_ursi/URSIGA08/papers/F06p5.pdf*. (Visited June 02 2009), Optical Crossing.

HEMMATI, H. (Ed.) (2006) *Deep space optical communications,* Canifornia, Wiley-Interscience

JEGANATHAN, M. & IONOV, P. Multi-gigabits-per-second optical wireless communications. http://www.freespaceoptic.com/WhitePapers/Jeganathan%20%20(Optical%20Crossing).pdf, (Visited May 20 2009), Optical Crossing.

KARP, S., GAGLIARDI, R. M., MORAN, S. E. & STOTTS, L. B. (1988) *Optical Channels: fibers, cluds, water and the atmosphere,* New York, Plenum Press.

KAZAURA, K., OMAE, K., SUZUKI, T., MATSUMOTO, M., MUTAFUNGWA, E., MURAKAMI, T., TAKAHASHI, K., MATSUMOTO, H., WAKAMORI, K. & ARIMOTO, Y. (2007) Performance evaluation of next generation free-space optical communication system. *IEICE Transaction of Electronics,* E90-C**,** 381-388.

KIASALEH, K. (2005) Performance of APD-based, PPM free-space optical communication systems in atmospheric turbulence. *IEEE Transactions on Communications,* 53**,** 1455-1461.

KILLINGER, D. (2002) Free space optics for laser communication through the air. *Optics & Photonics News***,** 36-42.

KIM, I. I. & KOREVAAR, E. (2001) Availability of free space optics and hybrid FSO/RF systems. *Proceedings of SPIE: Optical wireless communications IV,* 4530**,** 84-95.

OSCHE, G. R. (2002) *Optical Detection Theory for Laser Applications,* New Jersey, Wiley.

PELTON, J. N. (1998) Global satellite communications technology and systems. Baltimore, International Technology Research Institute, World Technology (WTEC) Division.

POPOOLA, W. O. & GHASSEMLOOY, Z. (2009) BPSK subcarrier intensity modulated free-space optical communications in atmospheric turbulence. *IEEE/OSA Journal of Lightwave Technology, ,* 27**,** 967-973.

POPOOLA, W. O., GHASSEMLOOY, Z., ALLEN, J. I. H., LEITGEB, E. & GAO, S. (2008) Free-space optical communication employing subcarrier modulation and spatial diversity in atmospheric turbulence channel. *IET Optoelectronic,* 2**,** 16-23.

PRATT, W. K. (1969) *Laser Communication Systems,* New York, John Wiley & Sons, Inc.

PROAKIS, J. G. (2004) *Digital Communications,* New York, McGraw-Hill.

ROCKWELL, D., A. & MECHERLE, S., G. (2001) Optical wireless: Low-cost, broadband, optical accesss. www.freespaceoptic.com/WhitePapers/optical_wireless.pdf.

WILLEBRAND, H. & GHUMAN, B. S. (2002) *Free Space Optics: Enabling optical Connectivity in today's network,* Indianapolis, IN, SAMS publishing.

Non Mechanical Compact Optical Transceiver for Wireless Communications with a VCSEL Array

Morio Toyoshima, Naoki Miyashita, Yoshihisa Takayama, Hiroo Kunimori
National Institute of Information and Communications Technology
Japan

Shinichi Kimura
Tokyo University of Science
Japan

1. Introduction

The advantages of optical communications as compared to radiowave (RF) communications include a wider bandwidth, a larger capacity, lower power consumption, more compact equipment, greater security against eavesdropping, better protection against interference, and the absence of regulatory restrictions (Hyde & Edelson, 1997). Moreover, the demand for high-data-rate transmission from spaceborne observation platforms is steadily increasing (Toyoshima, 2005a). Free-space optical (FSO) communications systems are expected to play an important role in providing such high-data-rate communications, and optical technologies for satellite networks are expected to revolutionize space system architecture (Chan, 2003). However, maintaining a line of sight between transceivers is particularly difficult because of the small divergence angle of the laser beams. Thus, minimizing the requirements for the tracking system and ensuring the steady operation of the onboard optical terminal are important for the realization of commercial applications.

Optical terminals tend to have a large mass because optical tracking systems need to have mechanically movable parts for coarse laser pointing and tracking. Reductions in the mass, power, and volume can decrease the interference to the other missions on the satellites. Non-mechanical movable architecture is very attractive for the robust and lifelong operation of the optical terminal in orbit. The small satellite community still uses 9.6-kbps communication links by employing ham radio communications because of resource constraints in the nano-class satellites (Nakaya et al., 2003; Miyashita et al., 2006). The compact terminal can be used in nano-class satellites that have a mass of the order of a few tens of kilograms. In addition, there is a significant advantage with regard to the frequency-licensing problems faced by satellites, and the optical frequency carrier will be very useful to the small satellite community.

In this chapter, a non-mechanical compact laser communications terminal is proposed for future applications. A laser beam is transmitted by selecting the laser pixel corresponding to the direction of the optical signal received from the counter terminal. The beams are not deflected by a mechanical mirror. Instead, they are turned on and off one after the other in accordance with the direction from which optical signals are received. Multiple-input multiple-output operation is possible in this configuration. The non-mechanical transceiver can facilitate a reduction in the size of the optical system. This chapter presents the basic concept of a non-mechanical compact transceiver with a two-dimensional laser array for optical wireless communications.

2. History of optical space communications

In Japan, the National Institute of Information and Communications Technology (NICT, formerly CRL) performed the first bidirectional laser communication by using the laser communications equipment (LCE) on Engineering Test Satellite VI (ETS VI), which was launched in August 1994. Due to the failure of its apogee engine, it was injected into an elliptical orbit, and it never reached the geostationary earth orbit (GEO). Nevertheless, both uplink and downlink transmissions at wavelengths of 0.514 and 0.830 μm, respectively, and a data rate of 1 Mbps were demonstrated (Arimoto et al., 1995). In 1995, the Jet Propulsion Laboratory (JPL) and NICT conducted a ground-to-space optical communications demonstration called the Ground-to-Orbit Lasercom Demonstration (GOLD) (Wilson et al., 1998).

The European Space Agency (ESA) conducted the Semiconductor Laser Intersatellite Link Experiment (SILEX) using laser communications terminals designed for 50-Mbps LEO-GEO intersatellite link (ISL) applications. A SILEX demonstration conducted in November 2001 that linked the SPOT-4 satellite to the Advanced Relay Technology Mission Satellite (ARTEMIS) marked the first transmission of an image at 50 Mbps over a laser link (Nielsen et al., 2002). A ground-to-ARTEMIS optical communications experiment was successfully conducted in November 2001 from the ESA's optical ground station in Tenerife, Spain (Reyes et al., 2003). In September 2003, optical acquisition, tracking, and communication tests between ARTEMIS in orbit and the Laser Utilizing Communications Equipment (LUCE) engineering model on the ground, which was developed for the Optical Inter-orbit Communications Engineering Test Satellite (OICETS, or "Kirari" in Japanese), were successfully performed. The compatibility of their optical interfaces had been tested before the launch of OICETS (Toyoshima et al., 2005b).

The acquisition and tracking of satellites are problematic because of their high angular velocity. In the Strategic Defense Initiative Organization's (SDIO) Relay Mirror Experiment (RME) in the early 1990s, three laser beams (two argon-ion laser beams at 0.488 and 0.514 μm and a Nd:YAG laser beam at 1.06 μm) were projected from two ground sites at the U.S. Air Force Maui Optical Station (AMOS) in Hawaii. These beams were retroreflected and reflected from the RME satellite at 350 km. A distortion in the uplinked beam intensity profile at the satellite was reported, but information on the optical communication was not mentioned (Lightsey et al., 1994). In 2000, the Ballistic Missile Defense Organization (BMDO) developed laser communications terminal equipment for the Space Technology Research Vehicle 2 (STRV2) experiment onboard a LEO satellite. However, due to a large attitude error in the host satellite, the experiment failed (Kim et al., 2001).

OICETS was launched by a Dnepr launch vehicle from the Baikonur cosmodrome in the Republic of Kazakhstan and placed in a LEO at 610 km with an inclination of 97.8º. The functioning of the satellite's systems was checked for the first three months, and during this time, stars and planets were acquired and tracked. In December 2005, the demonstration of the first bidirectional laser communication between OICETS and ARTEMIS was successfully conducted with a return link of 50 Mbps and a forward link of 2 Mbps (Jono et al., 2006; Takayama et al., 2007; Toyoshima et al., 2006; Toyoshima et al., 2007). Following these intersatellite experiments, the Kirari optical communication demonstration experiments were conducted by using the NICT optical ground station (KODEN) in collaboration with the Japan Aerospace Exploration Agency (JAXA) in March, May, and September 2006. Thus, the ground-to-OICETS laser communication experiment was the first in-orbit demonstration using a LEO satellite.

In March 2008, a 5.5-Gbps intersatellite optical communication link was successfully tested by using German laser terminals; this data rate is the fastest in space-borne communications to date. The laser terminals were tested in space for broadband data transmission between two satellites, the German satellite TerraSAR-X and the U.S. satellite NFIRE. Optical intersatellite links were established for a range of approximately 5000 km and were operated flawlessly at a data rate of 5.5 Gbps in both directions.

Through the research and development for FSO communications conducted in NICT, it is considered that a compact communications terminal will have applicability in FSO communications in the future. In this paper, we propose the concept of a compact free-space laser communications terminal with a vertical-cavity surface-emitting laser (VCSEL) array. There are no mechanically moving parts in this optical system. This compact terminal can receive optical communications signals from multiple platforms and can transmit multiple optical communications beams to the counter terminals. Therefore, such an optical system can serve as a multiple-input multiple-output (MIMO) system.

3. Conceptual terminal design

3.1 System Configuration

Figure 1 shows the configuration of the proposed compact laser communications transceiver. The laser beam from the counter terminal passes through the telescope lens, is reflected from the beam splitter and is detected by the CCD sensor. The CCD sensor detects the direction of the line of sight of the counter terminal, and one of the array lasers is selected according to the direction of the signal received by the CCD. A CCD with a pixel size equal to that of the XGA (1280 × 1024) is used. The centroid of the pixels is calculated in the computer, and the laser beam corresponding to the direction of the centroid is turned on. Figure 2 shows the photograph of the manufactured compact laser communications transceiver and the control computer system. With this configuration, multiple inputs from multiple platforms can possible with the parallel laser spot detection processing, and the MIMO configuration can be possible as well (Short et al., 1991).
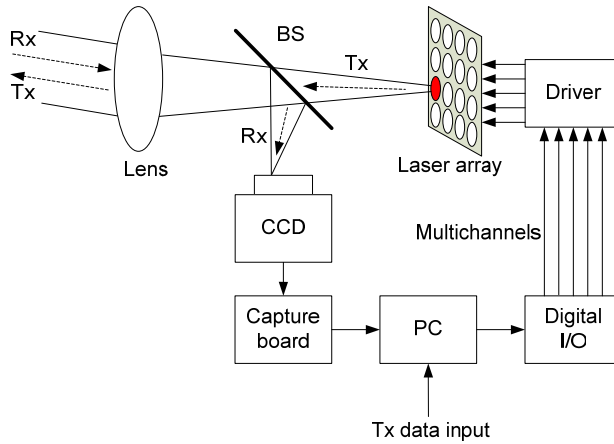
Fig. 1. Configuration of the proposed compact laser communications transceiver
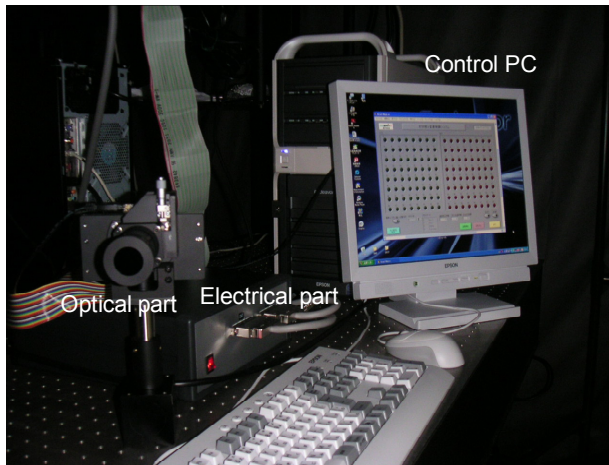

Fig. 2. Photograph of the manufactured compact laser communications transceiver

3.2 Optical part of the transceiver

The laser beam is transmitted from the two-dimensional laser array through the beam splitter and the telescope lens. The laser beam is selected by the centroid calculation in the computer. The beam divergence angle of the selected laser beam covers the angular interval between adjacent laser arrays (Cap et al., 2007). Two adjacent laser beams are turned on simultaneously to ensure that the laser transmission is not interrupted and to maintain a constant optical intensity at the counter terminal. Figure 3 shows the beam transmission configuration for a two-dimensional laser array. With this transmission method, the transmitted laser beam is not interrupted during the tracking of the counter terminal. Each laser beam is combined by an interval at the half width at half maximum (HWHM). Therefore, if the two adjacent laser beams are turned on simultaneously, the optical intensity can be almost constant at the counter terminal.

For the transmitter, we use an 8 × 8 VCSEL array, as shown in Figure 4, for the first evaluation model. We choose a VCSEL because it is easy to arrange it in an array, there is no mechanical part, and it is readily available. The maximum output power of one pixel is 4 mW at a wavelength of 850 nm as shown in Table 1. The laser diode can be modulated at above 2.5 GHz. An example of the beam pattern is measured by a beam profiler at 30 cm distance from the compact laser communications transceiver as shown in Figure 5. All the VCSELs could be turned on individually. The beam pattern from the VCSEL used here exhibits a donuts mode instead of the Gaussian one. The beam divergence for this evaluation model was designed to be 2 degrees for one VCSEL, which is more broadened beam divergence than that of the real system because the number of the VCSEL array is not enough. As shown in the next subsection, a large number of pixels are required depending on the angular coverage of the transmitter.

Parameter	Value
Number of the array	64 (8 × 8)
Maximum output power of one pixel	4 mW
Wavelength	850 nm
Beam divergence angle	20-30 degrees
Minimum frequency response	2.5 GHz

Table 1. Specification of the VCSEL array
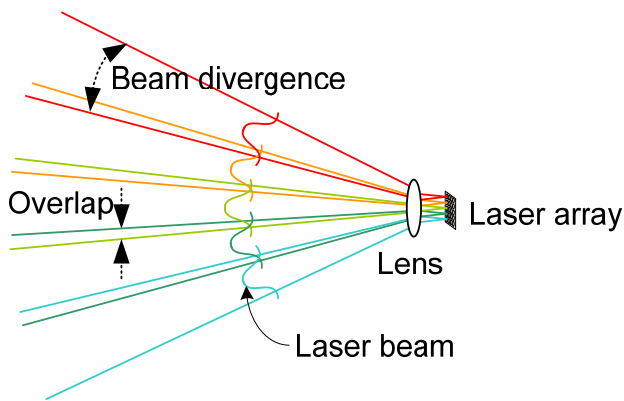
Fig. 3. Laser beam transmission method

Fig. 4. Photograph of the 8 × 8 VCSEL array

Fig. 5. Example of the beam pattern measured at 30 cm distance from the compact laser communication transceiver

Figure 6 shows the optical part of the manufactured compact laser communications transceiver. The small telescope consists of 9 lenses. The VSCEL is mounted at the end of the small telescope and the CCD sensor is mounted on the upper side of the telescope as shown

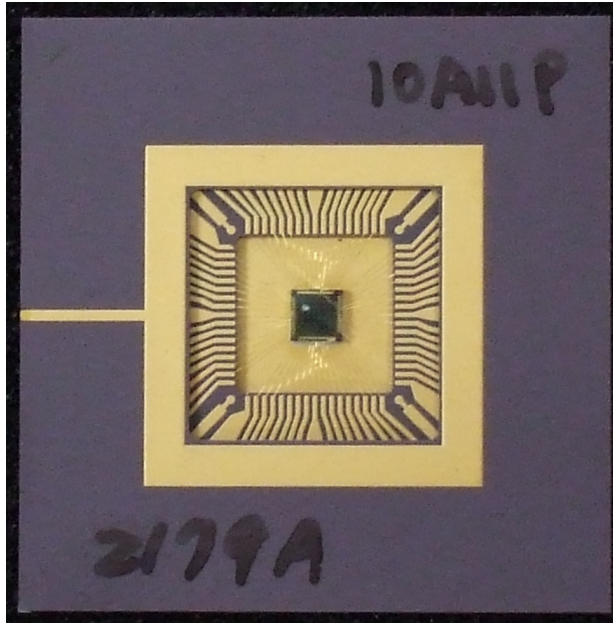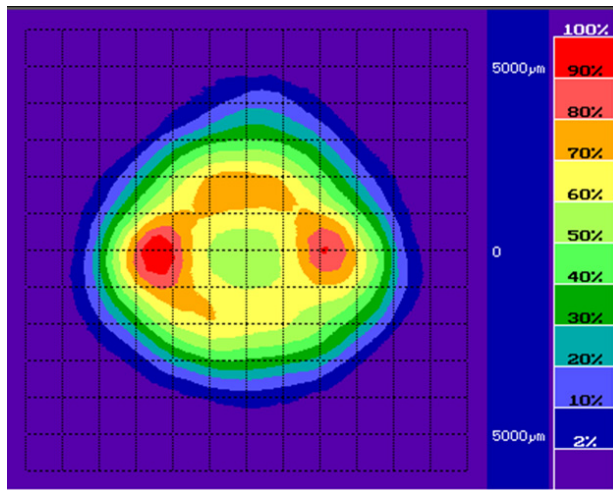in Figure 6. The size of the optical part of the telescope (lens mount) is 13.5 × 6 × 11 cm, and the mass is 1 kg as shown in Table 2.

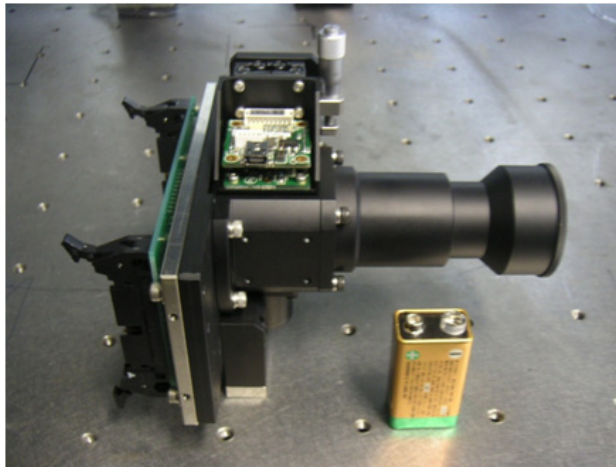Fig. 6. Optical part of the compact laser communication transceiver

Fig. 7. Electrical part of the compact laser communication transceiver

Resource		Value
Optical part	Mass	1 kg
	Size (lens mount)	15 × 12 × 12 cm (13.5 × 6 × 11 cm)
Electrical part	Mass	3.1 kg
	Size	27 × 26 × 10 cm
	Power	< 10 W

Table 2. Resources of the compact laser communication transceiver

3.3 Electrical part of the transceiver

Laser beams in the VCSEL array are modulated according to the received laser spot extracted by the control computer system as shown in Figure 1. Two 32-channel digital I/O boards are installed and can transmit the data at a data rate of 25 Mbps. Figure 7 shows the photograph of the electrical part of the manufactured laser driver. The electrical part as shown in Figure 7 can drive 64 channels of the VCSEL by the selected signal from the digital I/O boards. The laser diode is driven at an average power of 2 mW by the driver electronics. The electrical part of the compact laser communications transceiver has the mass of 3.1 kg, the size of 27 × 26 × 10 cm, and the power consumption of less than 10 W as shown in Table 2.

Parameters	Unit	Uplink	Downlink
TX power	W	1.00	0.002
	dBm	30.0	3.0
Pixel size of laser array	-	-	1024 × 1024
Beam diameter at telescope	cm	72.1	0.125
Pointing error	μrad	23.1	0.0
TX beam divergence (FWe²M)	μrad	1000.0	869.3
Angular coverage (FWHM)	degree	0.06	60.0
TX optic loss	dB	–2.0	–2.0
Wavelength	m	8.15E–07	8.50E–07
Average pointing loss	dB	0.0	0.0
TX gain	dB	72.0	76.3
Distance	m	1.00E+06	1.00E+06
Space loss	dB	–263.8	–263.4
Atmospheric transmission	dB	-6.0	–6.0
RX antenna diameter	cm	3.0	150.0
RX gain	dB	101.3	134.9
RX optic loss	dB	–2.0	–2.0
RX power	dBm	–70.5	–59.2
Data rate	bps	1.00E+05	1.00E+06
Sensitivity (@BER of 10^{-6})	Photons/bit	2200	2200
	dBm	–72.7	–62.9
Link margin	dB	2.2	3.6

Table 3. Link budget analysis of an optical communication link between a ground station and a satellite

3.4 System analysis

Table 3 summarizes the results of the link budget analysis for the proposed compact optical transceiver. The optical link is designed to connect an optical ground station to a LEO satellite. The output laser power for a pixel of the VCSEL array is assumed to be 2 mW at 850 nm wavelength. The beam divergence angle is set at 869 μrad for a single laser pixel for

the full width at $1/e^2$ maximum (FWe²M), and the angular coverage of the transmitter is 60° for a 1024 × 1024 array, which is enough to cover the attitude error of the satellites. The overlap of the beams is set to occur at the HWHM. The beam pointing error can be considered as zero because the transmitting power can be doubled by turning on the adjacent two VCSELs simultaneously. The beam diameter of the transmitter is 1.25 mm. The link distance is assumed to be 1000 km from the optical ground station to the LEO satellite. If we use a receiving telescope with a 1.5-m diameter, the link margin will be 3.6 dB at a data rate of 1 Mbps. The data rate is rather lower as compared with that of a trunk optical communications line; however, some nano-class satellites still use a data rate of the order of 1.2 k to 9.6 kbps (Nakaya et al., 2003; Miyashita et al., 2006). Despite the low data rate, this system is advantageous because it overcomes the frequency-licensing problem of the satellites. The compact optical transceiver can be used in such the nano-class satellite community for space communications in the future.

Table 4 tabulates the results of the link budget analysis for a horizontal terrestrial optical communication link on the ground. By using the proposed non-mechanical compact transceiver, a data rate of 1 Gbps can be realized for a distance up to 1 km with an angular coverage of 60°. The diameter of the receiver aperture is 5 cm in this case. If a receiver aperture with a diameter of 3 cm is used, a data rate of 1 Gbps can be realized for a distance up to 700 m. As shown in Table 4, the proposed non-mechanical method can be applied to terrestrial free-space laser communications. If the proposed terminal can be made very compact, mobile users can use the high-data-rate optical link without the mechanical tracking system on the ground, like a digital camera. Setting up the optical transceivers is easy and their installation is not complicated. In the future, the applicable fields for the optical transceivers will include not only satellite communications but also high-speed cell phone communications, wireless LAN, mobile communications, and building-to-building fixed high data rate communications without any difficulties.

4. Conclusion

A non-mechanical and very compact optical transceiver has been presented. A VCSEL array is used in the transceiver, and the laser pixel that is turned on depends on the direction of the counter terminal from which the CCD receives a signal. The mass, volume, and power of the proposed system can be reduced because it contains no mechanically movable structures. In this study, an 8 × 8 VCSEL has been used. The angular coverage of the optical terminal can be extended if a larger number of pixels are used. As there are no regulatory restrictions on the use of the optical frequency, the proposed compact laser communications transceiver will be useful for downloading data from nano-class satellites in future applications.

Parameters	Unit	Horizontal link
TX power	W	0.002
	dBm	3.0
Pixel size of laser array	-	1024 × 1024
Beam diameter at telescope	cm	0.125
Pointing error	μrad	0.0
TX beam divergence (FWe²M)	μrad	869.3
Angular coverage (FWHM)	degree	60.0
TX optic loss	dB	–2.0
Wavelength	m	8.50E–07
Average pointing loss	dB	0.0
TX gain	dB	76.3
Distance	m	1.00E+03
Space loss	dB	–203.4
Atmospheric transmission	dB	–6.0
RX antenna diameter	cm	5.0
RX gain	dB	105.3
RX optic loss	dB	–2.0
RX power	dBm	–28.8
Data rate	bps	1.00E+09
Sensitivity (@BER of 10^{-6})	photons/bit	2200
	dBm	–32.9
Link margin	dB	**4.1**

Table 4. Link budget analysis of a horizontal terrestrial laser communication link on the ground

5. References

Arimoto, Y.; Toyoshima, M., Toyoda, M., Takahashi, T., Shikatani, M. & Araki K. (1995). Preliminary result on laser communication experiment using Engineering Test Satellite-VI (ETS-VI), *Proc. SPIE*, Vol. 2381, pp. 151–158

Cap, G. A.; Refai, H. H. & Sluss, Jr., J. J. (2007). Omnidirectional free-space optical (FSO) receivers, *Proc. SPIE*, Vol. 6551-26, pp. 1–8

Chan, V. W. S. (2003). Optical satellite networks, *Journal of Lightwave Technology*, Vol. 21, pp. 2811–2827

Hyde, G. & Edelson, B. I. (1997). Laser satellite communications: Current status and directions, *Space Policy*, Vol. 13, pp. 47–54

Jono, T.; Takayama, Y., Kura, N., Ohinata, K., Koyama, Y., Shiratama, K., Sodnik, Z., Demelenne, B. Bird, A. & Arai, K. (2006) OICETS on-orbit laser communication experiments," *Proc. SPIE*, Vol. 6105, pp. 13–23

Kim, I. I.; Riley, B., Wong, N. M., Mitchell, M., Brown, W., Hakakha, H., Adhikari, P. & Korevaar, E. J. (2001). Lessons learned from the STRV-2 satellite-to-ground lasercom experiment, *Proc. SPIE*, Vol. 4272, pp. 1–15

Lightsey, P. A. (1994). Scintillation in ground-to-space and retroreflected laser beams," *Opt. Eng.*, Vol. 33, No. 8, pp. 2535–2543

Miyashita, N.; Konoue, K., Omagari, K., Imai, K., Yabe, H., Miyamoto, K., Iljic, T., Usuda, T., Fujiwara, K., Masumoto, S., Konda, Y., Sugita, S., Yamanaka T., & Matunaga, S. (2006). Ground operation and flight report of Pico-satellite Cute-1.7 + APD, *International Symposium on Space Technology and Science (25th ISTS)*

Nakaya, K.; Konoue, K., Sawada, H., Ui, K., Okada, H., Miyashita, N., Iai, M., Urabe, T., Yamaguchi, N., Kashiwa, M., Omagari, K., Morita, I. & Matunaga, S. (2003). Tokyo Tech CubeSat: CUTE-I -Design and development of flight model and future plan-, *21st AIAA International Communication Satellite System Conference & Exhibit*, Yokohama, vol. 2003-2388, April 15–19

Nielsen, T. T.; Oppenhaeuser, G., Laurent, B. & Planche, G. (2002). In-orbit test results of the optical intersatellite link, SILEX. A milestone in satellite communication, *Proceedings of the 53rd International Astronautical Congress*, Vol. IAC-02-M.2.01, pp. 1–11, Houston, October

Reyes, M.; Chueca, S., Alonso, A., Viera, T. & Sodnik, Z. (2003). Analysis of the preliminary optical links between ARTEMIS and the Optical Ground Station," *Proc. SPIE*, Vol. 4821, pp. 33–43

Short, R. C.; Cosgrove, M., Clark, D. L. & Oleski, P. (1991) Performance of a demonstration system for simultaneous laser beacon tracking and low data rate optical communications with multiple platforms, *Proc. SPIE*, Vol. 1417, pp. 464–475

Takayama, Y.; Jono, T., Toyoshima, M. Kunimori, H., Giggenbach, D., Perlot, N., Knapek, M., Shiratama, K., Abe, J., & Arai, K. (2007) Tracking and pointing characteristics of OICETS optical terminal in communication demonstrations with ground stations (Invited Paper), *Proc. SPIE*, Vol. 6457A, 6457A-06

Toyoshima, M. (2005a). Trends in satellite communications and the role of optical free-space communications [Invited], *Journal of Optical Networking*, Vol. 4, pp. 300–311

Toyoshima, M.; Yamakawa, S., Yamawaki, T., Arai, K., Reyes, M., Alonso, A., Sodnik, Z. & Demelenne, B. (2005b). Long-term statistics of laser beam propagation in an optical ground-to-geostationary satellite communications link," *IEEE Trans. Antennas and Propagat.*, Vol. 53, No. 2, pp. 842–850

Toyoshima, M.; Kunimori, H., Jono, T., Takayama, Y. & Arai, K. (2006) Measurement of atmospheric turbulence in a ground-to-low earth orbit optical link, *2006 Joint Conference on Satellite Communications (JC-SAT 2006)*, Vol. SAT2006-37, pp. 119–124, Jeju-do, Korea, October 20

Toyoshima, M.; Takahashi, T., Suzuki, K., Kimura, S., Takizawa, K., Kuri, T., Klaus, W., Toyoda, M., Kunimori, H., Jono, T., Takayama, Y. & Arai, K. (2007). Laser beam propagation in ground-to-OICETS laser communication experiments, *Proc. SPIE*, Vol. 6551, 6551-09

Wilson, K. E.; Lesh, J. R., Araki, K. & Arimoto Y. (1998). Overview of the Ground-to-Orbit Lasercom Demonstration, *Space Communications*, Vol. 15, pp. 89–95

Edited by Salma Ait Fares and Fumiyuki Adachi

Mobile and wireless communications applications have a clear impact on improving the humanity wellbeing. From cell phones to wireless internet to home and office devices, most of the applications are converted from wired into wireless communication. Smart and advanced wireless communication environments represent the future technology and evolutionary development step in homes, hospitals, industrial, vehicular and transportation systems. A very appealing research area in these environments has been the wireless ad hoc, sensor and mesh networks. These networks rely on ultra low powered processing nodes that sense surrounding environment temperature, pressure, humidity, motion or chemical hazards, etc. Moreover, the radio frequency (RF) transceiver nodes of such networks require the design of transmitter and receiver equipped with high performance building blocks including antennas, power and low noise amplifiers, mixers and voltage controlled oscillators. Nowadays, the researchers are facing several challenges to design such building blocks while complying with ultra low power consumption, small area and high performance constraints. CMOS technology represents an excellent candidate to facilitate the integration of the whole transceiver on a single chip. However, several challenges have to be tackled while designing and using nanoscale CMOS technologies and require innovative idea from researchers and circuits designers. While major researchers and applications have been focusing on RF wireless communication, optical wireless communication based system has started to draw some attention from researchers for a terrestrial system as well as for aerial and satellite terminals. This renewed interested in optical wireless communications is driven by several advantages such as no licensing requirements policy, no RF radiation hazards, and no need to dig up roads besides its large bandwidth and low power consumption. This second part of the book, Mobile and Wireless Communications: Key Technologies and Future Applications, covers the recent development in ad hoc and sensor networks, the implementation of state of the art of wireless transceivers building blocks and recent development on optical wireless communication systems. We hope that this book will be useful for students, researchers and practitioners in their research studies.

IntechOpen