RECENT ADVANCES IN WIRELESS COMMUNICATIONS AND NETWORKS

Edited by **Jia-Chin Lin**

INTECHWEB.ORG

Recent Advances in Wireless Communications and Networks
Edited by Jia-Chin Lin

Published by InTech
Janeza Trdine 9, 51000 Rijeka, Croatia

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published articles. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

Publishing Process Manager Niksa Mandic
Technical Editor Teodora Smiljanic
Cover Designer Jan Hyrat
Image Copyright Edyta Pawlowska, 2010. Used under license from Shutterstock.com

First published July, 2011
Printed in Croatia

A free online edition of this book is available at www.intechopen.com
Additional hard copies can be obtained from orders@intechweb.org

Recent Advances in Wireless Communications and Networks, Edited by Jia-Chin Lin
p. cm.
ISBN 978-953-307-274-6

Contents

Preface

Many exciting impacts on our daily life are shortly anticipated due to recent advances in wireless communication networks that enable real-time multimedia services to be provided via mobile broadband Internet on a wide variety of terminal devices. The trend is mainly driven by the evolution of wireless networks and advanced wireless information and communication technology (ICT). The progression to fourth-generation (4G) or IMT-advanced systems is expected to significantly change usage habits and introduce new services, such as the services supported by higher spectral-efficiency communication technology and self-configurable, high-feasibility networks. The Third Generation Partnership Project (3GPP) Long-Term Evolution (LTE) continues to be enhanced as this book is being written.

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

This book, titled Recent Advances in Wireless Communications and Networks, focuses on the current hottest issues from the lowest layers to the upper layers of wireless communication networks and provides “real-time” research progress on these issues. In my endeavor to edit this book, I have made every effort to ask the authors to systematically organize the information on these topics to make it easily accessible to readers of any level. The editor also maintains the balance between current research results and their theoretical support. In this book, a variety of novel techniques in wireless communications and networks are investigated. The authors attempt to present these topics in detail. Insightful and reader-friendly descriptions are presented to nourish readers of any level, from practicing and knowledgeable communication engineers to beginning or professional researchers. All interested readers can easily find noteworthy materials in much greater detail than in previous publications and in the references cited in these chapters.

This book is composed of twenty chapters that were authored by the most knowledgeable and successful researchers in the world. Each chapter was written in an introductory style beginning with the fundamentals, describing approaches to the hottest issues and concluding with a comprehensive discussion. The content in each chapter is taken from many publications in prestigious journals and followed by fruitful insights. The chapters in this book also provide many references for relevant topics, and interested readers will find these references helpful when they explore these topics further. These twenty chapters are arranged in order from the lowest layer to the upper layers of wireless communication. This book was naturally partitioned into 3 main parts. Part A consists of eight chapters that are devoted to physical layer (PHY) and medium access control (MAC) layer research. Part B consists of five chapters that are devoted to upper layer research. Finally, Part C consists of seven chapters that are devoted to applications and realizations.

Chapter 1 is an introduction to topics at an inner receiver in wireless communications, including a historical perspective and a description of Cramér-Rao-like bounds and relevant applications to estimation techniques in wireless communications. Chapter 2 conducts a thorough review of the initial synchronization techniques applied to wireless orthogonal-frequency-division-multiplexing (OFDM) communications. Chapter 3 is devoted to deeply investigating novel techniques of inter-carrier interference (ICI) reduction in practical OFDM communications. Chapter 4 is focused on multiple-antenna techniques from diversity, spatial multiplexing to beamforming techniques. Chapter 5 deeply investigates diversity management techniques in MIMO-OFDM communication systems. Chapter 6 thoroughly investigates resource allocation methods in OFDMA broadcast channels. This is the downlink scenario in either LTE-A or IEEE 802.16m wireless communications. Chapter 7 is dedicated to primary user detection in multi-antenna environments. Spectral sensing techniques are considered as the most important issue in recent research regarding cognitive radios (CRs) or cognitive networks. Chapter 8 focuses on multi-cell cooperation methodology. In third generation mobile communications, macro-diversity was investigated. In similar environments, multi-cell cooperation may be expected in opportunistic communication networks.

Chapter 9 covers a novel technique on joint call admission control in integrated wireless local area networks (WLAN) and cellular networks. Chapter 10 is devoted to studying near-optimal nonlinear forwarding strategies in two-hop MIMO relaying scenarios. Chapter 11 discusses a research on connectivity support in heterogeneous wireless networks for next-generation multimedia communication networks. Chapter 12 studies the use of a stream control transmission protocol (SCTP) in wireless communication networks. Chapter 13 investigates traffic control for composite wireless access route of IEEE802.11/16 links.

Chapter 14 comprehensively covers wireless sensor networks. Chapter 15 is a complete study of CRs built by a software-defined radio (SDR) platform. Chapter 16 deals with VoIP calls during rush hours in LTE communications. Chapter 17 demonstrates a semantics-based mobile web content transcoding framework. Chapter 18 proposes novel and effective power supply architectures for wireless communication systems. Chapter 19 applies wireless communication technology to smart structural technology. Chapter 20 provides extending applications of dielectric elastomer artificial muscles to wireless communication systems.

In summary, this book covers broad areas of communications and networks. The introductions, derivations, discussions and references in this book significantly improve the readers' understanding of communications and networks and encourage them to actively explore these broad, exciting and rapidly-evolving research areas.

Jia-Chin Lin
Distinguished Professor
Dept. of Commun. Engr.
National Central University,
Taiwan, R.O.C.

Part 1

Physcial and MAC Layers

1

A Study of Cramér-Rao-Like Bounds and Their Applications to Wireless Communications

Kao-Peng Chou and Jia-Chin Lin
Department of Communication Engineering, National Central University
Jhongli, Taoyuan,
Taiwan

1. Introduction

Estimation theory has been developed over centuries. There are several approaches to utilizing this theory; in this chapter, these approaches are classified into three types. Type I includes the oldest two methods, the least squares (LS) and moment methods; both of these methods are non-optimal estimators. The least squares method was introduced by Carl Friedrich Gauss. Least squares problems fall into linear and non-linear categories. The linear least squares problem is also known as regression analysis in statistics, which have a closed form solution. An important feature of the least squares method is that no probabilistic assumptions of the data are made. Therefore, the linear least squares approach is used for parameter estimation, especially for low complexity design (Lin, 2008; 2009). The design goal of the least squares estimator is to find a linear function of observations whose expectation is a linear function of the unknown parameter with minimum variance. In addition, the least squares method corresponds to the maximum likelihood (ML) criterion if the experimental errors are normally distributed and can also be derived from the moment estimation. As an alternative to the LS method, the moment method is another simple parameter estimation method with probabilistic assumptions of the data. The general moment method was introduced by K. Pearson. The main procedure in the moment method involves equating the unknown parameter to a moment of distribution, then replacing the moment with a sample moment to obtain the moment estimator. Although the moment estimator has no optimal properties, the accuracy can be validated through lengthy data measurements. This is mainly because the estimator based on moment can be maintained to be consistent. Type II includes the methods of minimum variance unbiased estimator (MVUE) and the Bayesian approach, which are both optimal in terms of possible minimum estimation error, i.e., statistical efficiency. MVUE is the best guess of an unknown parameter. The standard MVUE procedure includes two steps. In the first step, the Cramer-Rao lower bound is determined, and the ability of some estimator to approach the bound. In the second step, the Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem is applied. The MVUE can be produced by these two steps. Moreover, a linear MVUE might be found under more restricted conditions.

In the Bayesian method, the Bayesian philosophy begins with the cost function, and the expected cost with respect to the parameter is the risk. The design goal of Bayesian philosophy is to find an estimator that minimizes the average risk (Bayes risk). The most common cost function is a quadratic function because it measures the performance of the estimator in terms of the square of the estimation error. In this case, the Bayes risk is the mean square error (MSE), and thus, the Bayes estimate is a minimum mean square error (MMSE) estimator. Another common cost function is the absolute function, which regards the absolute estimate error as the Bayes risk. In this case, the Bayes estimate is a minimum mean absolute error (MMAE) estimator. Another estimation, which is not a proper Bayes estimation but fits within Bayes philosophy, is the maximum a posteriori (MAP) estimation. The MAP criterion considers the uniform cost function, and the parameter is discretely, randomly distributed under this assumption. Although this estimate usually only approximates the Bayes estimate for uniform cost, the MAP criterion is widely used for estimator design. Type III includes the maximum likelihood (ML) estimate, which is the most important estimation theory in the 20th century. The ML estimate can be referred to as an alternative MAP without knowledge of apriori probability of the parameters. The ML estimator is the most popular approach for obtaining a practical estimator, which was previously used by Gauss. The general method of estimation was first introduced by R. A. Fisher with the concepts of consistency, efficiency and sufficiency of the estimation function. The ML estimator is required when MVUE does not exist or cannot be found. An advantage of the ML estimator is that a practical estimation is easy to obtain through the prescribed procedures. Another advantage of this approach is that MVUE can be approximated due to its efficiency. Thus, from the theoretical and practical perspectives, the ML approach is the most important and widely used estimation method of this century (Lin, 2003).

Because the ML estimator is essential in estimation theory, the analysis of its performance is a benchmark of estimator design. This benchmark is commonly known as the Cramer-Rao lower bound (CRLB), which is named after Harald Cramer and Calyampudi Radhakrishna Rao. In section 2, the definition of the CRLB is introduced with several examples. A general case of CRLB under two common communication channels is then introduced in section 3. To establish basic knowledge of hybrid parameter estimation, random parameter estimation is presented in section 4. In section 5, Cramer-Rao-like bounds for hybrid parameter estimation are introduced and compared with each other. Lastly, we summarize some practical cases and compare these cases with modified CRB which is most common used Cramer-Rao-like bounds.

2. Cramer-Rao lower bound (CRLB)

The Cramer-Rao lower bound (CRLB) is a lower bound on the variance of any unbiased estimator. Many other variance bounds exist, but the CRLB is the easiest one to derive and is thus widely used in many estimation studies. This theory provides a benchmark for examining the performance of novel estimation algorithms and also highlights the impossibility of finding an unbiased estimator with a variance less than this lower bound.

Before introducing the definition of CRLB, there is a simple estimation example that may could help promote understanding of the basic CRLB concept.

Example 2.1

There is a simple signal transmission model with a transmitted signal s, a received signal $r[n]$ and an additive white Gaussian noise $w[n]$.

$$r[n] = s + w[n] \tag{1}$$

Here, the index n refers to the n'th observation. In this problem, the transmitted signal s is assumed to be an unknown parameter that is deterministic during n observations. The first idea estimate s takes one observation as our estimation, e.g., the n'th observation, namely $\hat{s} = r[n]$. To analyze the estimation accuracy, we check the likelihood function of $r[n]$ as shown.

$$p(r[n];s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \left(r[n] - s \right)^2 \right] \tag{2}$$

Substituting the estimator we chose in this likelihood function yields

$$p(\hat{s};s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \left(\hat{s} - s \right)^2 \right] \tag{3}$$

Now, the mean value is the target parameter s, and the estimation variance is σ^2. The estimation accuracy can then be determined as

$$\operatorname{var}(\hat{s}) = \sigma^2 = -E\left(\frac{\partial^2 \ln p(r[n];s)}{\partial s^2} \right)^{-1}. \tag{4}$$

Furthermore, we are interested in finding a more accurate estimator by lowering the variance σ^2. This can be achieved by exploiting multiple observations. Assuming the observation samples are identical independently distributed, the likelihood function for multiple observations is

$$p(\mathbf{r}[n];s) = \frac{1}{\left(2\pi\sigma^2 \right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N-1} \left(r[n] - s \right)^2 \right]. \tag{5}$$

A ML estimator can be derived in the same way as for a single observation to yield

$$\hat{s} = \frac{\sum_{n=1}^{N-1} r[n]}{N}, \tag{6}$$

which is an unbiased estimator, namely $E\{\hat{s}\} = s$. We can also find the estimation variance using equation (4); the result is similar to the single observation MLs with a factor N in the denominator:

$$\operatorname{var}(\hat{s}) = \frac{\sigma^2}{N}. \tag{7}$$

An extreme case occurs when N approaches ∞, and the process reduces the estimation variance to 0. From this simple example, we can summarize that the ultimate goal of estimator design is to find the minimum variance unbiased estimator (MVUE), and if we wish to illustrate the performance of our estimator, then estimation variance can be found through the likelihood function. Now, we are ready to define the CRLB (Kay, 1998).

<Theorem>

Assume the pdf, $p(r;\theta)$ *,satisfies the regularity condition*

$$E_{r;\theta}\left[\frac{\partial \ln p(r;\theta)}{\partial \theta}\right]=0 \quad \textit{for all } \theta . \tag{8}$$

Then, the variance of any unbiased estimator $\hat{\theta}$ *has a lower limitation*

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{-E_{r;\theta}\left[\dfrac{\partial^2 \ln p(r;\theta)}{\partial \theta^2}\right]} . \tag{9}$$

An unbiased estimator may be found that attains the bound for all θ *if and only if*

$$\frac{\partial \ln p(r;\theta)}{\partial \theta}=I(\theta)(g(r)-\theta) \tag{10}$$

for some function $I(\theta)$ *and* $g(r)$ *. This estimator can be stated as* $\hat{\theta}=g(r)$ *, which is a MVUE with variance* $1/I(\theta)$ *. To attain the variance lower bound, Fisher's information is defined as*

$$I(\theta)=-E_{r;\theta}\left[\frac{\partial^2 \ln p(r;\theta)}{\partial \theta^2}\right], \tag{11}$$

which is used to calculate the covariance matrices associated with maximum-likelihood estimates.
An unbiased estimator that achieves the variance lower bound is referred to as "efficient". In other words, an unbiased estimator that achieves the CRLB is an efficient estimator and must be MVUE. Figures 1 and 2 are illustrations of the relationship between a MVU estimator and the CRLB.

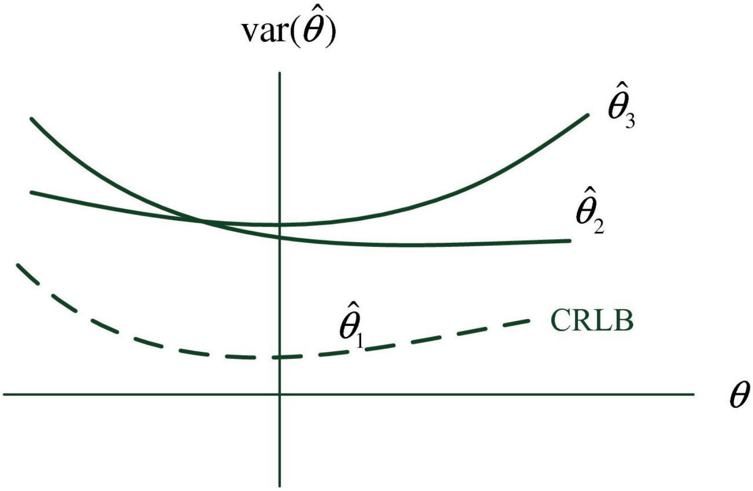


Fig. 1. $\hat{\theta}_1$ MVU and efficient

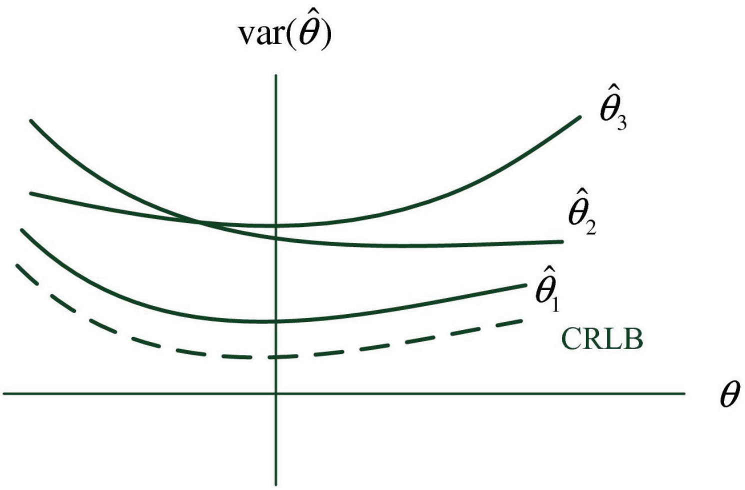


Fig. 2. $\hat{\theta}_1$ MVU and not efficient

Although there are some theories capable of finding MVUE by sufficient statistics and the Rao-Blackwell-Lehmann-Scheffe theorem, we will not introduce the details in this chapter. However, we encourage readers to fully inform themselves concering MVUE from the references in this chapter (Kay, 1998).

A question may be raised concerning why the minimum variance estimator should be an unbiased one. Although the unbiased estimator seems to sucessfully find an perfect estimator φ because the expectation value approaches the true parameter i.e., $E[\hat{\theta}] = \theta_0$, but a biased estimator may outperform than an unbiased one. For example, in some situations, the relationship between a MVUE and a Bayesian MSE estimator may be illustrated in figure 3.

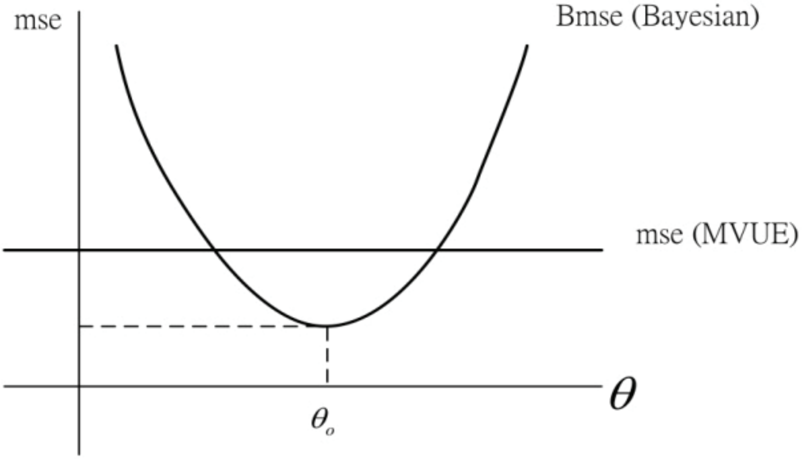


Fig. 3. MVUE vs. Bayesian estimator

In this example, the Bayesian MSE estimator is an unbiased estimator. The performance comparison in figure 3 shows that within a certain parameter interval, the biased Bayesian estimator may have lower estimation variance than MVUE's. However, this comparison also shows that the biased estimator performs terribly outside this interval. Thus, the unbiased estimator has an advantage in terms of consistent performance.

2.1 Asymptotic CRLB

For some cases in which the closed form of the CRLB may not be derived, the asymptotic CRLB can be used instead; this form can be attained by assuming that infinite observation samples are available. Under this assumption, we have an observation sample with an infinite signal-to-noise ratio (SNR).

3. General case CRLB

3.1 Gaussian noise

The AWGN channel is the most common channel model in wireless communication, which was also used in the example in the last section. In example 2.1, we only consider the estimate of symbol s. Now, a general form of any parameter θ is derived.

Example 3.1

Assuming symbol s is transmitted with a general unknown parameter θ and added with an AWGN $w_n(t)$. The signal model is describe as

$$r_n(t) = s(t;\theta) + w_n(t)\,, \tag{12}$$

where n indicate the nth observation. Following the general CRLB derivation steps, the likelihood function is found first and differentiation with respect to θ is then performed twice.

$$p(r_n(t);s(t),\theta) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N-1}\left[r_n(t)-s(t;\theta)\right]^2\right] \tag{13}$$

$$\frac{\partial \ln p(r_n(t);s(t),\theta)}{\partial\theta} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left[r_n(t)-s(t;\theta)\right]\frac{\partial s(t;\theta)}{\partial\theta} \tag{14}$$

$$\frac{\partial^2 \ln p(r_n(t);s(t),\theta)}{\partial\theta^2} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s(t;\theta)}{\partial\theta}\right)^2 + \left[r_n(t)-s(t;\theta)\right]\frac{\partial^2 s(t;\theta)}{\partial\theta^2} \tag{15}$$

Taking the expectation of $\frac{\partial^2 \ln p(r_n(t);s(t),\theta)}{\partial\theta^2}$ with respect to $p(r;s,\theta)$ into Fisher's information yields

$$I(\theta) = -E_{r;s,\theta}\left\{\frac{\partial^2 \ln p(r_n(t);s(t),\theta)}{\partial\theta^2}\right\} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s(t;\theta)}{\partial\theta}\right)^2 \tag{16}$$

Finally, the inverse recipocal of the Fisher's information produced by the CRLB in the AWGN channel.

$$\text{var}(\hat{\theta}) \geq \frac{1}{I(\theta)} = \frac{\sigma^2}{\sum_{n=0}^{N-1}\left(\frac{\partial s(t;\theta)}{\partial \theta}\right)^2} \tag{17}$$

3.2 Complex Gaussian channel

Another commonly seen channel is complex Gaussian channel. The mobile communication and wireless communication usually introduce the Rayleigh fading due to multipath delay spread and Doppler shift. In numerical simulation we may use the Jake's (Clarke) model, but in theoretical analysis, complex Gaussian channel is more popular, because it has Rayleigh distributed amplitude with an uniformly distributed phase, which is convenient to use and without loss of generality.

Example 3.2

The signal model can be extended from the general AWGN channel model. We multiply the Rayleigh distributed channel gain α_0 and the uniformly distributed channel phase $e^{-j\phi_0}$ with the symbol $s(t;\theta_u)$.

$$r_n(t) = \alpha_0 e^{-j\phi_0} s(t;\theta_u) + w_n(t) \tag{18}$$

Alternatively, using complex coordinates, i.e., the Gaussian distributed α_I and α_Q with mean η_A and variance σ_A^2 yields

$$r_n(t) = (\alpha_I + j\alpha_Q)s(t;\theta) + w_n(t) \tag{19}$$

Because the α_I , α_Q and $w_n(t)$ terms are Gaussian distributed, the received signal $r_n(t)$ is also Gaussian distributed. To find the joint likelihood function, the mean m_r and variance σ_r^2 of the received signal should be derived.

$$m_r = \eta_A(1+j)s(t;\theta) \tag{20}$$

$$\sigma_r^2 = 2\sigma_A^2 P_s(t;\theta) + 2\sigma_N^2 \tag{21}$$

Here, $P_s(t;\theta) = s(t;\theta)s(t;\theta)^*$ is the power of the transmitted signal. The joint likelihood function turns out is then described by

$$p_r(r(t);s(t;\theta)) = \frac{1}{\sqrt{2\pi\sigma_r^2}}\exp(-\frac{(r(t)-m_r)^2}{2\sigma_r^2}) \tag{22}$$

4. Random parameter estimation

In previous sections, some basic knowledge of estimation bounds were introduced based on unknown parameters with random interference. These kinds of estimation problems are categorized in the classical estimation approach. Some properties of estimation methods are listed in Table 1.

	Parameter types	Sample distribution	Parameter distribution
LS	Unknown	Unknown	Non
Moment	Unknown	Known	Non
MVUE	Unknown	Known	Non
Bayesian	Random	Known	Known
MAP	Random	Known	Known
ML	Both	Known	Uniform

Table 1. Some estimation properties

Another research area focuses on random parameters estimation, and several approaches, including the Bayesian theorem, MAP and ML, are widely used already. One of the most popular and well-known Bayesian approache is the MMSE estimator. Below, the MMSE will be briefly introduced with an example.

Example 4.1

Assuming that we received signal $r(t)$ that is composed of a random symbol s and white Gaussian noise $w(t)$, the following relationship can be described.

$$r(t) = s + w(t) \tag{23}$$

The conditional pdf of $r(t)$ with a priori information can be stated as

$$p(r(t);s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \sum_{n=0}^{N-1} (r_n(t) - s)^2 \right) \tag{24}$$

Using Bayes' rule,

$$p(r(t);s) = \frac{p(s;r(t))p(r(t))}{p(s)} \tag{25}$$

After certain computations, the conditional pdf with a posteriori information is obtained as

$$p(s;r(t)) = \frac{1}{\sqrt{2\pi\sigma_{s;r}^2}} \exp\left(-\frac{1}{2\sigma_{s;r}^2} (s - \mu_{s;r})^2 \right), \tag{26}$$

where

$$\sigma_{s;r}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_s^2}}; \tag{27}$$

$$\mu_{s;r} = \left(\frac{N}{\sigma^2} \overline{x} + \frac{\mu_s}{\sigma_s^2} \right) \sigma_{s;r}^2 . \tag{28}$$

The MMSE estimator is then determined as

$$\hat{s} = E\{s \mid r(t)\} = \mu_{s;r} = \alpha \overline{x} + (1 - \alpha)\mu_s \tag{29}$$

where

$$\alpha = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma^2}{N}} \tag{30}$$

The Bayesian mean square error is defined as

$$Bmse(\hat{s}) = E[(s-\hat{s})^2] = \frac{\sigma^2}{N}\left(\frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma^2}{N}}\right) \le \frac{\sigma^2}{N} \tag{31}$$

As $\sigma_s^2 \to \infty$ i.e., without any information from a prior knowledge, the bound would be the same with the sample mean estimator. This result can be compared with that of the first example in this chapter, and an important concept of Bayesian estimator is revealed: any prior knowledge will result in higher accuracy of the Bayesian estimator.

5. Hybrid parameter estimation

In addition to classical estimation and random parameter estimation, there is a more complicated scenario called hybrid parameter estimation. In hybrid parameter estimation, the desired parameter is a vector that is composed of several unknown paramters and random parameters. The parameter vector can be constructed as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_r^T & \boldsymbol{\theta}_u^T \end{bmatrix}^T, \tag{32}$$

where $\boldsymbol{\theta}_r$ is a random parameter vector and $\boldsymbol{\theta}_u$ is an unknown parameter vector. Because we are considering the random parameters, we assume that we have some prior knowledge of these parameters, such as the probability distribution function. Several techniques for calculating hybrid parameter Cramer-Rao like bounds are described below.

5.1 CRLB with nuisance parameter

In our first case, $\boldsymbol{\theta}_r$ is treated as a nuisance parameter, which means that these random parameter are undesired.

Example 5.1

Reformulating the signal model and likelihood function yields

$$r_n(t) = s(t;\boldsymbol{\theta}) + w_n(t) \tag{33}$$

$$p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{[r_n(t)-s(t;\boldsymbol{\theta})]^2}{2\sigma^2}). \tag{34}$$

Because we assumed that the pdf is well-known and these denoted parameters are unimportant, the marginal likelihood function is derived first, and the nuisance parameters are integrated out of the equation.

$$p(r_n(t);s(t),\boldsymbol{\theta}_u) = \int_{\theta_r} p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)p(\boldsymbol{\theta}_r)d\boldsymbol{\theta}_r \tag{35}$$

Now, the resultant problem becomes a classical estimation problem, and the CRLB can be derived step by step.

1. $$\frac{\partial \ln p(r_n(t);s(t,\boldsymbol{\theta}_u))}{\partial \boldsymbol{\theta}_u} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left[r_n(t)-s(t,\boldsymbol{\theta}_u)\right]\frac{\partial s(t,\boldsymbol{\theta}_u)}{\partial \boldsymbol{\theta}_u} \tag{36}$$

2. $$\frac{\partial^2 \ln p(r_n(t);s(t,\boldsymbol{\theta}_u))}{\partial {\boldsymbol{\theta}_u}^2} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s(t,\boldsymbol{\theta}_u)}{\partial \boldsymbol{\theta}_u}\right)^2 + \left[r_n(t)-s(t,\boldsymbol{\theta}_u)\right]\frac{\partial^2 s(t,\boldsymbol{\theta}_u)}{\partial {\boldsymbol{\theta}_u}^2} \tag{37}$$

3. $$I(\boldsymbol{\theta})_{i,j} = E_r\left\{\frac{\partial \ln p(r_n(t);s(t,\boldsymbol{\theta}_u))}{\partial \theta_i}\frac{\partial \ln p(r_n(t);s(t,\boldsymbol{\theta}_u))}{\partial \theta_j}\right\} \tag{38}$$

4. $$CRLB(\hat{\theta}_i) = \left[\frac{1}{I(\boldsymbol{\theta})}\right]_{i,i} \le \mathrm{var}(\hat{\theta}_i) \tag{39}$$

5.2 Hybrid CRLB

In some scenarios, the effect of these ramdom parameters cannot be ignored. Another method that considers the joint pdf called joint estimation. The CRLB for this kind of joint estimation is called hybrid Cramer-Rao bound (HCRB). The derivation process is nearly identical to that of ordinary CRLB; the likelihood function is determined, and partial differentiation with respect to the desired parameter is performed twice.

$$r_n(t) = s(t;\boldsymbol{\theta}) + w_n(t) \tag{40}$$

$$p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left[r_n(t)-s(t;\boldsymbol{\theta})\right]^2}{2\sigma^2}\right) \tag{41}$$

$$\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left[r_n(t)-s(t;\boldsymbol{\theta})\right]\frac{\partial s(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{42}$$

$$\frac{\partial^2 \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial \boldsymbol{\theta}^2} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial s(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2 + \left[r_n(t)-s(t;\boldsymbol{\theta})\right]\frac{\partial^2 s(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \tag{43}$$

Because the joint pdf is considered, the expection of Fisher's information should be taken with respect to $p(r(t),\theta_r)$

$$I(\boldsymbol{\theta})_{i,j} = E_{r,\boldsymbol{\theta}_r}\left\{\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial \theta_i}\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial \theta_j}\right\} \tag{44}$$

The joint pdf $p(r(t),\theta_r)$ is not easy to determine, and an alternative approach using double layer expectation which computes the expectation with respect to the conditional pdf first. We define the information matrix with respect to the conditional pdf $p(r(t);\theta_r)$ as

$$I(\boldsymbol{\theta}_0)_{i,j} = E_{r;\boldsymbol{\theta}_r}\left\{\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial\theta_i}\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial\theta_j}|\boldsymbol{\theta}_r\right\}. \tag{45}$$

Then expectation is computed with respect to $p(\theta_r)$, and all of the random pararameters are eliminated.

$$\begin{aligned} I(\boldsymbol{\theta})_{i,j} &= E_{\boldsymbol{\theta}_r}\left\{E_{r;\boldsymbol{\theta}_r}\left\{\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial\theta_i}\frac{\partial \ln p(r_n(t),\boldsymbol{\theta}_r;s(t),\boldsymbol{\theta}_u)}{\partial\theta_j}|\boldsymbol{\theta}_r\right\}\right\} \\ &= E_{\boldsymbol{\theta}_r}\left\{I(\boldsymbol{\theta}_0)_{i,j}\right\} \end{aligned} \tag{46}$$

Finally, the HCRB is derived as

$$HCRB(\hat{\theta}_i) = \left[\frac{1}{I(\boldsymbol{\theta})}\right]_{i,i} \le \operatorname{var}(\hat{\theta}_i)\,. \tag{47}$$

5.3 Modified CRLB

During the process of deriving the HCRB, an important step involves taking the inverse of the Fisher's information matrix. In some cases, the inverse of the Fisher's information matrix may not exist or cannot be derived into a closed form lower bound. We can then try the modified or simplified bound, such as the MCRB. Instead of taking the inverse of the matrix first, we select the desired estimation element from the information matrix first and then execute the inverse step. After choosing the desired estimation element, the Fisher's information is no longer in a matrix form, and derivation is easier.

$$MCRB(\hat{\theta}_i) = \left[\frac{1}{I(\boldsymbol{\theta})_{i,i}}\right] \le \operatorname{var}(\hat{\theta}_i) \tag{48}$$

An previously reported example can help distinguish the difference between these CR-like bounds (F. Gini, 2000).

Example 5.2

When considering a data-aided joint frequency offset estimation case, the signal model can be described as

$$r_n(t) = Ae^{-j2\pi f_D t}s(t) + w_n(t) \tag{49}$$

Here, A is the complex channel, which can be rewritten as $A = \alpha_0 e^{-j\phi_0} = \alpha_I + j\alpha_Q$, and $e^{-j2\pi f_D t}$ represents the frequency offset. The estimation parameter matrix $\boldsymbol{\theta} = [f_D \quad \alpha_I \quad \alpha_Q]^T$ can be defined. Because this is a data-aided case, $s(t)$ can be a pilot or preamble, and we can assume that $s(t)s(t)^* = 1$ without loss of generality. Then the signal after pilot removal is

$$\begin{aligned} x_n(t) &= r_n(t)s(t)^* \\ &= (\alpha_I + j\alpha_Q)e^{j2\pi f_D t} + v_n(t) \end{aligned} \tag{50}$$

$x_n(t)$ is also Gaussian distributed. Following the derivation of S. M. Kay (1998) and F. Gini (2000), we can find the conditional Fisher's information matrix.

$$I(\theta_0)=\begin{bmatrix} \frac{2\pi^2 N(N-1)(2N-1)}{3\sigma_N^2}(\alpha_I^2+\alpha_Q^2) & -\frac{\pi N(N-1)}{\sigma_N^2}\alpha_Q & \frac{\pi N(N-1)}{\sigma_N^2}\alpha_I \\ -\frac{\pi N(N-1)}{\sigma_N^2}\alpha_Q & \frac{N}{\sigma_N^2}+\frac{(\alpha_I-\eta_A)^2}{\sigma_A^2} & 0 \\ \frac{\pi N(N-1)}{\sigma_N^2}\alpha_I & 0 & \frac{N}{\sigma_N^2}+\frac{(\alpha_Q-\eta_A)^2}{\sigma_A^2} \end{bmatrix} \tag{51}$$

By computing the expectation of α , the Fisher's information for the frequency offset is

$$I(f_D)=E_\alpha\{I(\theta_0)\} \tag{52}$$

Then the MCRB is derived as

$$MCRB(f_D)=\frac{1}{\left[I(f_D)\right]_{11}}=\frac{3}{4\pi^2 N(N-1)(2N-1)\rho} \tag{53}$$

where $\rho=(\eta_A^2+\sigma_A^2)/\sigma_N^2$ is the SNR. Now, the difference between the MCRB and the HCRB can be checked. As mentioned previously, the HCRB is

$$\begin{aligned} HCRB(f_D)&=\left[\frac{1}{I(f_D)}\right]_{11} \\ &=\frac{3(K_R+1)(K_R+1+N\rho)}{2\pi^2 N(N-1)}\frac{1}{2(2N-1)(K_R+1)(K_R+1+N\rho)-3N(N-1)\rho K_R} \end{aligned}, \tag{54}$$

where $K_R=\eta_A^2/\sigma_A^2$ is the Rice factor, which is the power ratio between direct path signal and other scatter path signals. A comparison of the HCRB and MCRB can be evaluted as.

$$\frac{HCRB(f_D)}{MCRB(f_D)}=\frac{2(2N-1)(K_R+1)(K_R+1+N\rho)}{2(2N-1)(K_R+1)(K_R+1+N\rho)-3N(N-1)\rho K_R}. \tag{55}$$

Based on the equation above, in the general case, the ratio is always larger than 1, which means that the HCRB is generally a tighter bound than the MCRB. Conversely, when $K_R\to 0$ or $K_R\to\infty$, the ratio of HCRB to MCRB approaches 1. It is interesting that these two bounds only meet for two extreme scenarios, namely the Rayleigh channel and direct path.

5.4 Miller Chang bound

The Miller Chang bound (MCB) is proposed by R. W. Miller and C. B. Chang (1978). They state that the MCB can apply to a more restricted class of estimator that is unbiased for each value of the nuisance parameter, which is referred to as locally unbiased, whereas the standard Cramer-Rao bound (CRB) can applies to any estimators that are unbiased over the ensemble. The Miller Chang bound is defined as

$$MCB(\hat{\theta}_i) = E_{\boldsymbol{\theta}_r}\left\{\frac{1}{I(\boldsymbol{\theta}_0)_{i,i}}\right\}. \tag{56}$$

The MCB has a similar form to the MCRB, but the MCB is always tighter than the MCRB. More directly, the MCB applies to more restricted estimators than the CRLB, which implies that the MCB is tighter than CRB, and the MCRB is looser than the CRB, which was derived by A. N. D'Andrea (1994). Therefore, the MCB is tighter than the MCRB. Alternatively, we can also explain this relationship using Jensen's inequality for any convex function φ and random variable x

$$\varphi\left(E[x]\right) \le E\left[\varphi(x)\right]. \tag{57}$$

In our case, the inverse function for a positive defined matrix is a convex function, so

$$MCRB(\hat{\theta}_i) = \frac{1}{E_{\boldsymbol{\theta}_r}\left\{I(\boldsymbol{\theta}_0)_{i,i}\right\}} \le E_{\boldsymbol{\theta}_r}\left\{\frac{1}{I(\boldsymbol{\theta}_0)_{i,i}}\right\} = MCB(\hat{\theta}_i). \tag{58}$$

Now, from example 5.2 in the MCRB subsection, the MCB of the joint estimated frequency offset is

$$\begin{aligned} MCB(f_D) &= \frac{3\sigma_N^2}{2\pi^2 N(N-1)(2N-1)} E_\alpha\left\{\frac{1}{\alpha_I^2+\alpha_Q^2}\right\} \\ &= MCRB(f_D) E_\alpha\left\{\frac{2(\eta_A^2+\sigma_A^2)}{\alpha_I^2+\alpha_Q^2}\right\} \end{aligned} \tag{59}$$

The final result still remains the expectation term, so it cannot be derived into a closed form. Although the MCB is a tighter bound than the MCRB, the MCRB is more likely to derive into a closed form. In addition, the MCB requires a locally unbiased estimator, which is also a harsh restriction for estimator design, so the MCRB is more popular for theoretical analysis.

5.5 Summary of the relationship between Cramer-Rao-like bounds

Some of the relationship between Cramer-Rao-like bounds has been derived previously (Reuven, 1997). In this work, they consider the signal model with Gaussian distributed channel gain and an unknown timing delay. We can also derive this relationship from our examples in subsection 5. Following from example 5.1, if we carry through the calculation to the end, then we will obtain the marginal CRB of the frequency offset f_D.

$$CRB(f_D) = \frac{3(K_R+1+N\rho)}{2\pi^2 N(N-1)\rho[N(N+1)\rho+2(2N-1)K_R]} \tag{60}$$

Then, this result is compared with that for the HCRB, which was derived in equation (55).

$$CRB(f_D) = HCRB(f_D)\frac{2(2N-1)(K_R+1)(K_R+1+N\rho)-3N(N-1)\rho K_R}{(K_R+1)[N(N+1)\rho+2(2N-1)K_R]} \tag{61}$$

After calculations, the CRB can be summarized into the HCRB multiplied by a function. We simplified the fraction in equation (62) and found that it is larger than 1 only if $N < 1$. This result implies that $CRB(f_D) \geq HCRB(f_D)$, and the relationship $HCRB(f_D) \geq MCRB(f_D)$ has been proven by equation (56). Another way to prove this is to use a corollary.

"For any positive defined matrix M, $\left[M^{-1}\right]_{11} \geq \left[M_{11}\right]^{-1}$, an equal occur if M is diagonal".

Finally, we summarize the relationship between CRB, HCRB and MCRB as

$$CRB(f_D) \geq HCRB(f_D) \geq MCRB(f_D) \tag{62}$$

However, the relationship between the MCB and MCRB was also derived in equations (58-59) using Jensen's inequality. Because the MCRB seems to be a looser bound in the Cramer-Rao-like bounds family, we normalized all other bounds to the MCRB, as shown in figure 4.

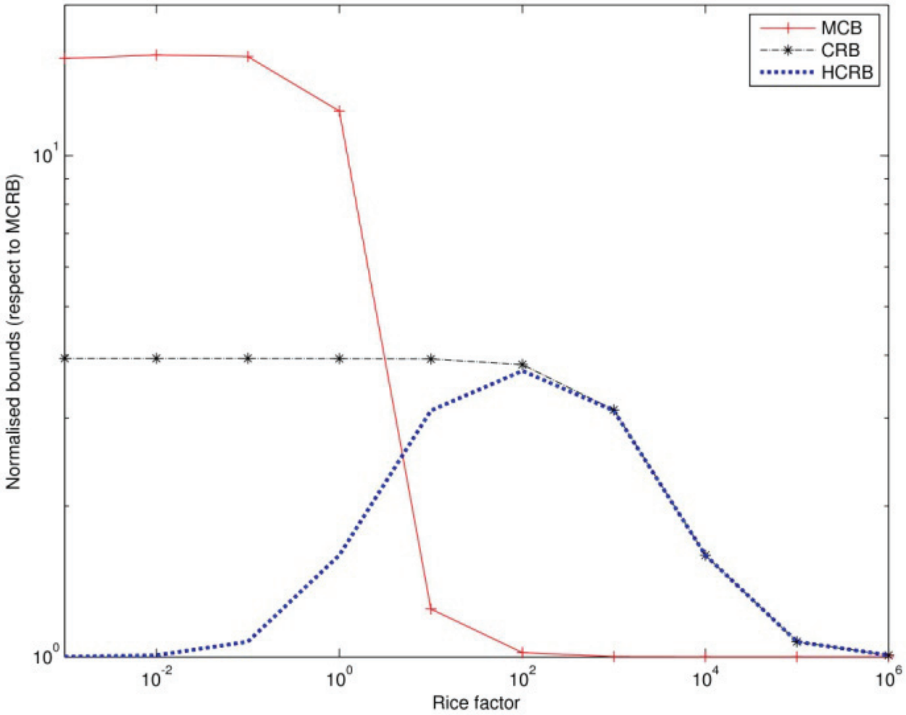


Fig. 4. Normalized bounds versus the Rice factor

From the figure above, the MCB exhibits drastic variation near $K_R = 1$, which indicates that the locally unbiased estimatior of f_D is difficult to find when the power of scatter signal is larger than line-of-sight (LOS) signal. Moreover, when $K_R = 0$ (Rayleigh channel), the ratio of the normalized MCB approaches infinity, which means that no locally unbiased estimator exists. The Rayleigh fading channel is the most frequently used channel model in a wireless communication environment, which is another reason that the MCRB is more popular than the MCB. In multiple parameters estimation, joint estimation techniques have been a popular topics recently. In terms of hybrid parameter joint estimation, the benchmark for comparison with is the HCRB. Based on equation (56) and figure 3, the HCRB has a feature that approaches the MCRB when $K_R \to 0$ or $K_R \to \infty$. As mentioned previously, the scenario $K_R \to 0$ implies the Rayleigh channel. The analysis shows that the MCRB is quite sufficient as a benchmark to design an estimator in the Rayleigh channel environment.

Some prior research has been reported on the relationship among the joint estimate initial phase, timing delay and frequency offset (D'Andrea , 1994). The author summarized and derived some cases in which the CRB is equal to the MCRB.

i. Estimation of ϕ when f_D, τ and data are known
ii. Estimation of τ when f_D, ϕ, and data are known
iii. Estimation of f_D with M-PSK modulation, when τ and differential data are available but ϕ is unknown.

Here, ϕ, f_D and τ are the initial phase , frequency offset and timing delay. Other cases may exist in which the CRB is equal to the MCRB, but these cases are difficult to analyze. An important conclusion here is that if an estimator approaches the MCRB, then the MCRB must be closed to the CRB.

6. Advanced topics

6.1 Carrier phase and clock recovery

As summarized by A. N. D'Andrea (1994), there are several synchronization techniques that can attain or approach the MCRB for a carrier phase θ and timing τ estimation. Under the assumption that the frequency offset and timing are known, $MCRB(\theta)$ can be attained using two algorithms.

i. Maximum likelihood decision-directed (ML-DD), proposed in H. Kobayashi (1971)
ii. Ad hoc non-data-aided (ad hoc NDA) method, proposed by A. J. Viterbi (1983).

The $MCRB(\tau)$ can also be attained using the ML-DD algorithm with derivative-matched filters (DMFs); however, the use of DMFs also makes the estimator impractical to implement. Several alternative algorithma have been found that can approach $MCRB(\tau)$ without using DMFs.

i. DD early-late scheme with $T / 2$ sample space, proposed by T. Jesupret (1991).
ii. DD scheme, proposed by K. H. Mueller (1976).
iii. NDA scheme, proposed by F. M. Gardner (1986).

Although these alternative algorithms can approach $MCRB(\tau)$ without using DMFs, they are subject to some restrictions that require θ to be known and a roll-off factor α that should be small.

6.2 Frequency offset estimation

In this subsection, three practical carrier frequency estimation techniques are overviewed and compared with the popular MCRB.

A. NDA loop algorithm

The first algorithm is a non-data-aided carrier frequency estimation; a block diagram representing this algorithm is shown in figure 5. The received signal $r(t)$ first passes through the matched filter $G^{*}(f)$ and the so-called "frequency-matched filter" $dG^{*}(f)/df$. Assuming that the timing is perfectly synchronized, the frequency error is described as

$$e_k = \mathrm{Re}\{x_k y_k^{*}\}\,. \tag{63}$$

Then, the frequency error passes through a loop filter and triggers the voltage-control oscillator (VCO) to compensate for the frequency offset. If the loop filter is implemented by a simple digital integrator, then the VCO output can be written as

$$\hat{f}_D(k+1) = \hat{f}_D(k) + \gamma e_k \tag{64}$$

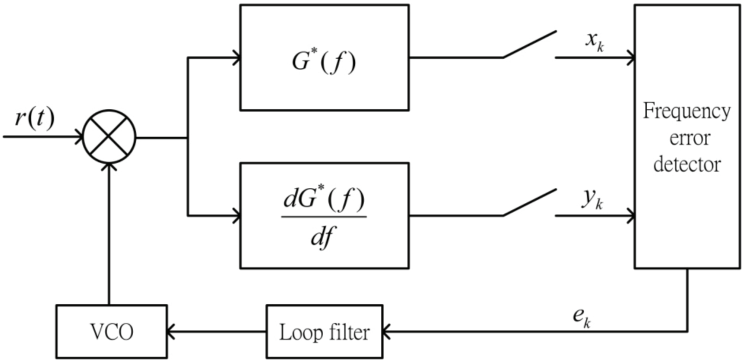


Fig. 5. NDA loop algorithm

The next step is to evaluate the estimation noise performance. There are three assumptions

i. The frequency errors are small as compared to the symbol rate.
ii. The pulse shaping filter $G^{*}(f)$ is a root-raised cosine function with a roll-off factor α .
iii. There is perfect timing delay synchronization

Under these assumptions, the frequency jitter is minimized, and the estimation variance of f_D is derived to be

$$\sigma_{f_D}^2 = \frac{4\alpha B_L T}{\pi^2 T^2} \frac{1}{E_s/N_0}\left(1 + \frac{1}{E_s/N_0}\right), \tag{65}$$

where B_L is the loop noise bandwidth and T is the symbol duration.

B. Differential decision-directed algorithm

The second algorithm is a differential decision-directed (DDD) algorithm that is used on PSK signals; the block diagram for this algorithm is shown in figure 5. This algorithm is similar to the NDA algorothm except for the frequency error generator. The assumptions for this algorithm include the following:

i. The frequency errors are small compared to the symbol rate.
ii. $G^{*}(f)$ is the same as was defined previously.
iii. Timing is perfectly synchronized.

Because we are discussing the M-PSK signal, we can denote our symbol by

$$c_k = \exp(j\varphi_k)\,, \tag{66}$$

where

$$\varphi_k = 2\pi n / M,\ \ n = 1,2...M\,. \tag{67}$$

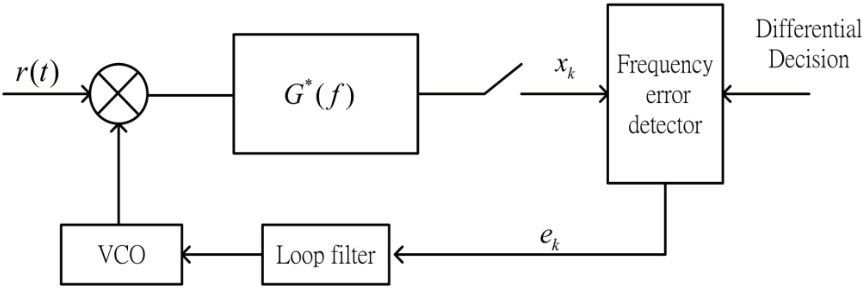


Fig. 6. Differential decision-directed (DDD) algorithm

Then the phase difference between x_k and x_{k-1} will be

$$\Delta\phi_k = \Delta\varphi_k + (f_D - \hat{f}_D)T + \delta_k\,, \tag{68}$$

where $\Delta\varphi_k$ is due to modulation, $(f_D - \hat{f}_D)$ is caused by estimation error, and δ_k is the phase noise with other interferences, which can be modeled as an uniformly distributed random variable from $-\pi$ to π. When the difference between x_k and x_{k-1} is correct, perfect $\Delta\hat{\varphi}_k = \Delta\varphi_k$ is obtained. The most important component of this block diagram is the frequency error that is defined as

$$e_k = \mathrm{Im}\left\{x_k x_{k-1}^* \exp(-j\Delta\hat{\varphi}_k)\right\} \tag{69}$$

The performance of the estimator is then

$$\sigma_{f_D}^2 = \frac{B_L T}{\pi^2 T^2}\frac{1}{E_s/N_0}\left(2B_L T + \frac{1}{E_s/N_0}\right) \tag{70}$$

Prior to the simulation, we assume that $B_L T = 5\times 10^{-3}$ and the QPSK signals have a roll-off factor $\alpha = 0.5$. The result is compared with the MCRB in figure 7.
As shown in figure 7, these two algorithms yield much greater variance than the MCRB, which indicates that there is still room for improvement.

C. Feed-forward NDA

The third algorithm is the feed-forward NDA for M-PSK signal modulation; the block diagram for this algorithm is shown in figure 8. The F function in the middle of the block diagram is a 4th-powered non-linear function. Similar to the previous analyses , the received phase can be separated into three parts:

i. A step-wise increasing quantity $2\pi M f_D kT$ due to the frequency error f_D.
ii. A constant initial phase.

iii. A phase noise caused by thermal noise and inter-symbol interference that is uniformly distributed from $-\pi$ to π .

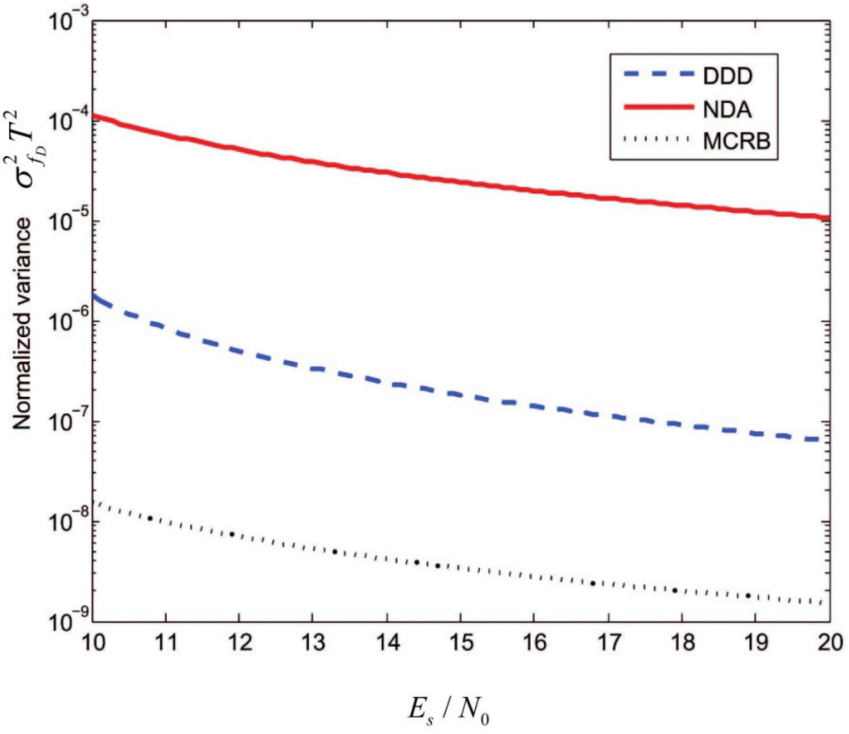


Fig. 7. Comparison of the variance of the two algorithms with that of the MCRB

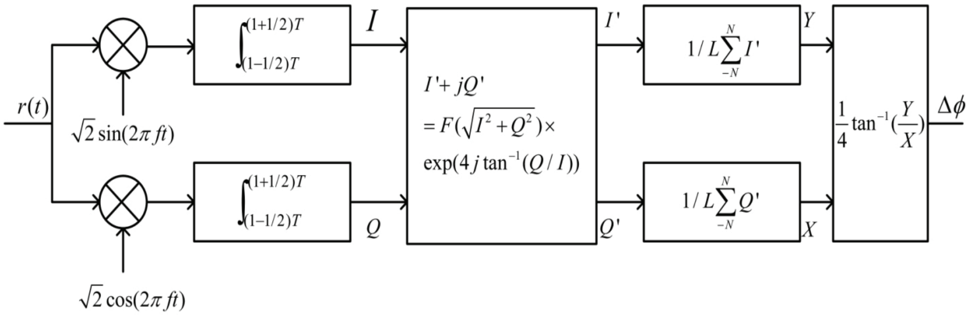


Fig. 8. Feed-forward NDA

The estimation variance has been derived (Bellini, 1990) in a scenario with a very high SNR, the estimation variance can be approached as

$$\sigma_{f_D}^2 \approx \frac{3}{2\pi^2 T^2 L(L^2\text{-}1)m} \frac{1}{E_s/N_0} \tag{71}$$

The MCRB in this case is

$$MCRB(f_D) = \frac{3T}{2\pi^3 (LT)^3} \frac{1}{E_s/N_0} \tag{72}$$

Thus, when $L \gg 1$ and $m = 1$,the algorithm performance will attain the MCRB. However, this result is obtained under very high SNR. Further research is needed to design estimators that can approach or attain the estimation bounds with less restriction.

7. References

Bellimi, S., Molinari, C. and Tartara, G. (1990). Digital Frequency Estimation in Burst Mode QPSK Transmission, *IEEE Trans. Commun.*, Vol.38, No.7 , (July 1990), pp. 959-961, ISSN: 0090-6778

Cramer, H. (1946). *Mathematical Method of Statistics*, Princeton University Press, ISBN-13: 978-0691005478, Uppsala, Sweden.

D'Andrea, A. N., Mengali, U. and Reggiannini, R. (1994). The Modified Cramer-Rao Bound and Its Application to Synchronization Problems, *IEEE Trans. Commun.*, Vol.42, No.2/3/4, (Febuary 1994), pp. 1391-1399, ISSN: 0090-6778

Gini, F. and Reggiannini, R. (2000). On the Use of Cramer-Rao-Like Bounds in the Presence of Random Nuisance Parameters, *IEEE Trans. Commun.*, Vol.48, No.12, (December 2000), pp. 2120-2126, ISSN 0090-6778.

Gardner, F. M. (1986). A BPSK/QPSK Timing Error Detecor for Samples Receivers, *IEEE Trans. Commun.*, Vol.34, No.5, (May 1986), pp. 423-429, ISSN: 0090-6778

Jesupret, T., Moeneclaey, M. and Ascheid, G. (1991). Digital Demodulator Synchronization, ESA Draft Final Report, ESTEC No. 8437-89-NL-RE., (Febuary 1991)

Kay, S. M. (1998). *Fundamentals of Statistical Signal Processing,* Prentice Hall, ISBN 0-13-345711-7, Upper Saddle River, New Jersey

Kobayashi, H. (1971). Simultaneous Adaptive Estimation and Decision Algorithm for Carrier Modulated Data Transmission Systems, *IEEE Trans. Commun.*, Vol.19, No.3, (June 1971), pp. 268-280, ISSN: 0018-9332

Kotz, S. and Johnson, N. L. (1993). *Breakthroughs in Statistics: Volume 1: Foundations and Basic Theory*, Springer-Verlag, ISBN: 0387940375, New York.

Lin, J. C. (2003). Maximum-Likelihood Frame Timing Instant and Frequency Offset Estimation for OFDM Communication Over A Fast Rayleigh Fading Channel, *IEEE Trans. Vehic. Technol.,* Vol.52, No.4, (July 2003), pp. 1049-1062.

Lin, J. C. (2008). Least-Squares Channel Estimation for Mobile OFDM Communication on Time-Varying Frequency-Selective Fading Channels, *IEEE Trans. Vehic. Technol.,* Vol.57, No.6, (November 2008), pp. 3538-3550.

Lin, J. C. (2009). Least-Squares Channel Estimation Assisted by Self-Interference Cancellation for Mobile PRP-OFDM Applications, *IET Commun.,* Vol.3, Iss.12, (December 2009), pp. 1907-1918.

Mueller, K. H. and Muller, M. (1976). Timing Recovery in Digital Synchronous Data Receivers, *IEEE Trans. Commun.*, Vol.24, No.5, (May 1976), pp. 516-530, ISSN: 0090-6778.

Miller, R. W. and Chang, C. B. (1978). A Modified Cramer-Rao Bound and its Applications, *IEEE Trans. On Inform. Throey,* Vol.IT-24, No.3, (May 1978), pp-389-400, ISSN : 0018-9448

Poor, H. V. (1994). *An Introduction to Signal Detection and Estimation*, Springer-Verlag, ISBN: 0-387-94173-8, New York.

Viterbi, A. J. and Viterbi, A. M. (1983). Nonlinear Estimation of PSK-Modulated Carrier Phase with Application to Burst Digital Transmission, *IEEE Trans. Inform. Throey,* Vol.IT-29, No.3, (July 1983), pp. 543-551, ISSN : 0018-9448.

2

Synchronization for OFDM-Based Systems

Yu-Ting Sun and Jia-Chin Lin
*National Central University, Taiwan,
R.O.C*

1. Introduction

Recently, orthogonal frequency division multiplexing (OFDM) techniques have received great interest in wireless communications for their high speed data transmission. OFDM improves robustness against narrowband interference or severely frequency-selective channel fades caused by long multipath delay spreads and impulsive noise. A single fade or interferer can cause the whole link to fail in a single carrier system. However, only a small portion of the subcarriers are damaged in a multicarrier system. In a classical frequency division multiplexing and parallel data systems, the signal frequency band is split into N nonoverlapping frequency subchannels that are each modulated with a corresponding individual symbol to eliminate interchannel interference. Nevertheless, available bandwidth utilization is too low to waste precious resources on conventional frequency division multiplexing systems. The OFDM technique with overlapping and orthogonal subchannels was proposed to increase spectrum efficiency. A high-rate serial signal stream is divided into many low-rate parallel streams; each parallel stream modulates a mutually orthogonal subchannel individually. Therefore, OFDM technologies have recently been chosen as candidates for fourth-generation (4G) mobile communications in a variety of standards, such as 802.16m and LTE/LTE-A.

2. OFDM fundamentals

2.1 System descriptions

The block diagram of an OFDM transceiver is shown in Fig. 1. Information bits are grouped and mapped using M-phase shift keying (MPSK) or quadrature amplitude modulation (QAM). Because an OFDM symbol consists of a sum of subcarriers, the $n-$th $N \times 1$ mapped signal symbol X_n is fed into the modulator using the inverse fast Fourier transform (IFFT). Then, the modulated signal x_n can be written as

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}, \quad n = 0,1,\cdots,N-1 \tag{1}$$

where N is the number of subcarriers or the IFFT size, k is the subcarrier index, n is the time index, and $1/N$ is the normalized frequency separation of the subcarriers. Note that x_n and X_k form an $N-$point discrete Fourier transform (DFT) pair. The relationship can be expressed as

$$X_n = \mathrm{DFT}_N\{x_n\} = \frac{1}{N}\sum_{k=0}^{N-1} x_k e^{-j2\pi kn/N}, \quad n = 0,1,\cdots,N-1 \tag{2}$$

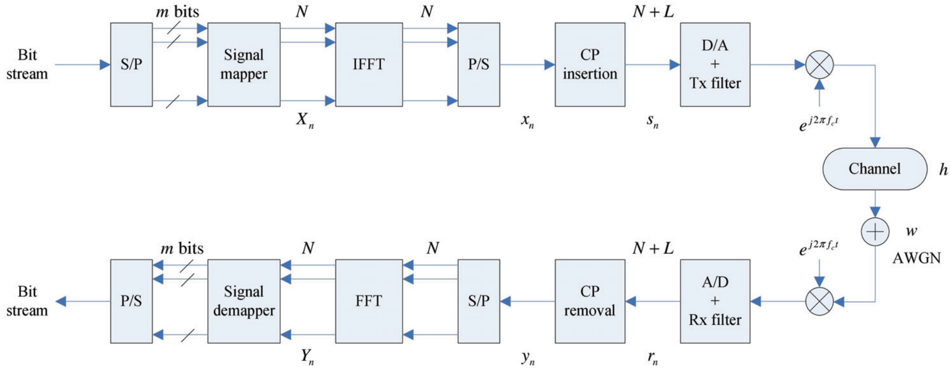


Fig. 1. The block diagram of the OFDM transceiver

The data symbol X_k can be recovered approximately by using a DFT operation at the receiver if the orthogonality of the OFDM symbol is not destroyed by intersymbol interference (ISI) and intercarrier interference (ICI). A cyclic prefix (CP) is used in an OFDM system to prevent ISI and ICI. The CP usually repeats the last L samples of an OFDM block and then is arranged in front of the block. The resulting symbol s_n can be represented as

$$s_n = \begin{cases} x_{N+n}, & n = -L,-L+1,\cdots,-1 \\ x_n, & n = 0,1,\cdots,N-1 \end{cases} \tag{3}$$

The transmitted signal may pass through a channel h depending on the environments. The receiver signal r_n can be written as

$$r_n = s_n \otimes h + w \tag{4}$$

where w denotes the additive white Gaussian noise (AWGN). The data symbol Y_n can be recovered by using a DFT operation and is determined as

$$Y_n = \frac{1}{N}\sum_{k=0}^{N-1} y_k e^{j2\pi kn/N}, \quad n = 0,1,\cdots,N-1 \tag{5}$$

Fig. 2 (a) shows the spectrum of an OFDM subchannel, and (b) shows an entire OFDM signal. At the maximum value of each subcarrier frequency, all other subcarrier spectra are null. The relationship between the OFDM block and CP is depicted clearly in Fig. 3.

The OFDM technique offers reliable effective transmission; however, it is far more vulnerable to symbol timing error and carrier frequency offset. Sensitivity to symbol timing offset is much higher in multicarrier communications than in single carrier communications because of intersymbol interference. The mismatch or instability of the local oscillator inevitably causes an offset in the carrier frequency that can cause a high bit error rate and performance degradation because of intercarrier interference. Therefore, the unknown

OFDM symbol arrival times and mismatch/instability of the oscillators in the transmitter and the receiver are two significant synchronization problems in the design of OFDM communications. A detailed description of symbol timing error and carrier frequency offset is given in the following sections.

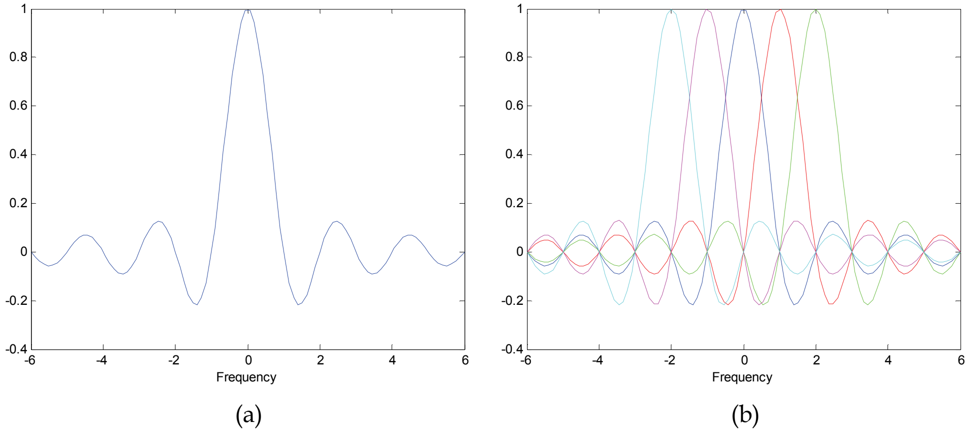


Fig. 2. Spectra of (a) an OFDM subchannel and (b) an OFDM signal

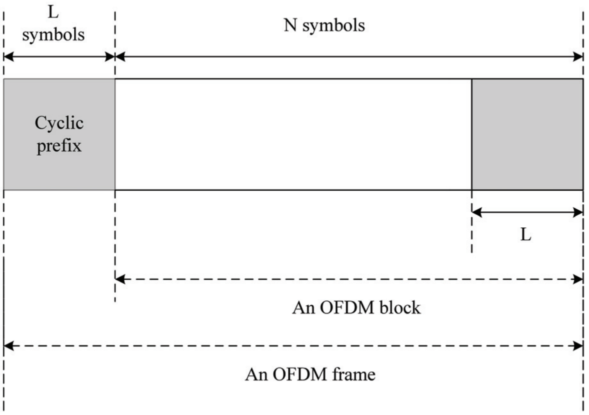


Fig. 3. An OFDM symbol with a cyclic prefix

2.2 Synchronization issues

2.2.1 Timing offset

OFDM systems exploit their unique features by using a guard interval with a cyclic prefix to eliminate intersymbol interference and intercarrier interference. In general, the symbol timing offset may vary in an interval that is equal to the guard time and does not cause intersymbol interference or intercarrier interference. OFDM systems have more robustness to compare with carrier frequency offset. However, a problem arises when the sampling frequency does not sample an accurate position; the sensitivity to symbol timing offset increases in OFDM systems. Receivers have to be tracked time-varying symbol timing offset, which results in time-varying phase changes. Intercarrier interference comes into being another attached problem. Because an error in the sampling frequency means an error in the FFT interval duration, the sampled subcarriers are no longer mutually orthogonal. The deviation is more severe as the delay spread in multipath fading increases; then, the tolerance for the delay spread is less than the expected value. As a result, timing synchronization in OFDM systems is an important design issue to minimize the loss of robustness.

2.2.2 Carrier frequency offset

In section 2.1, it is evident that at all OFDM subcarriers are orthogonal to each other when they have a different integer number of cycles in the FFT interval. The number of cycles is not an integer in FFT interval when a frequency offset exists. This phenomenon leads to intercarrier interference after the FFT. The output of FFT for each subcarrier contains an interfering term with interference power that is inversely proportional to the frequency spacing from all other subcarriers (Nee & Prasad, 2000). The amount of intercarrier interference for subcarriers in the middle of the OFDM spectrum is roughly twice as larger as that at the OFDM band edges because there are more interferers from interfering subcarriers on both sides. In practice, frequency-selective fading from the Doppler effect and/or mismatch and instability of the local oscillators in the transmitter and receiver cause carrier frequency offset. This effect invariably results in severe performance degradation in OFDM communications and leads to a high bit error rate. OFDM systems are more sensitive to carrier frequency offset; therefore, compensating frequency errors are very important.

3. Application scenarios

The major objectives for OFDM synchronization include identifying the beginning of individual OFDM symbol timing and ensuring the orthogonality of each subcarrier. Various algorithms have been proposed to estimate symbol timing and carrier frequency offset. These methods can be classified into two categories: data-aided algorithms and non-data-aided (also called blind) algorithms. By using known training sequences or pilot symbols, a data-aided algorithm can achieve high estimation accuracy and construct the structure simply. Data-aided algorithms require additional data blocks to transmit known synchronization information. Nevertheless, this method diminishes the efficiency of transmission to offer the possibility for synchronization. Non-data-aided (blind) algorithms were proposed to solve the inefficiency problem of the data-aided algorithm. Alternative techniques are based on the cyclic extension that is provided in OFDM communication systems. These techniques can achieve high spectrum efficiency but are more complicated.

In the data-aided technique, several synchronization symbols are directly inserted between the transmitted OFDM blocks; then, these pilot symbols are collected at the receiving end to extract frame timing information. However, the use of pilot symbols inevitably decreases the capacity and/or throughput of the overall system, thus making them suitable only in a startup/training mode. The data- aided technique can provide effectively synchronization with very high accuracy. Thus, it can be used to find coarse timing and frequency offset in the initial communication link. Several data-aided techniques have been proposed (Classen & Meyr, 1994, Daffara & Chouly, 1993, Kapoor et al., 1998, Luise & Reggiannini, 1996, Moose, 1994, Warner & Leung, 1993). Moreover, the SNR at the front end in the receiver is often too low to ineffectively detect pilot symbols; thus, a blind approach is usually much more desirable. A non-data-aided technique can adjust the fine timing and frequency after the preamble signal. Some non-data-aided techniques have been proposed (Bolcskei, 2001, Daffara & Adami, 1995, Lv et al., 2005, Okada et al., 1996, Park et al., 2004, Van de Beek et al., 1997).

3.1 Non-data-aided method

The cyclic extension has good correlation properties because the initial T_{CP} seconds of each symbol are the same as the final seconds in OFDM communications. The cyclic prefix is used to evaluate the autocorrelation with a lag of T . When a peak is found in the correlator output, the common estimates of the symbol timing and the frequency offset can be evaluated jointly. The correlation output can be expressed as

$$x(t) = \int_0^{T_{CP}} r(t-\tau) r^*(t-\tau-T) d\tau \tag{6}$$

where $r(t)$ is the received OFDM signal, $x(t)$ is the correlator output, τ denotes the timing offset. The correlator output can be utilized to estimate the carrier frequency offset when the symbol timing is found. The phase drift between T seconds is equivalent to the phase of the correlator output. Therefore, the carrier frequency offset can be estimated easily by dividing the correlator phase by $2\pi T$. The carrier frequency offset denotes the frequency offset normalized by the subcarrier spacing. Fig. 4 shows the block diagram of the correlator.

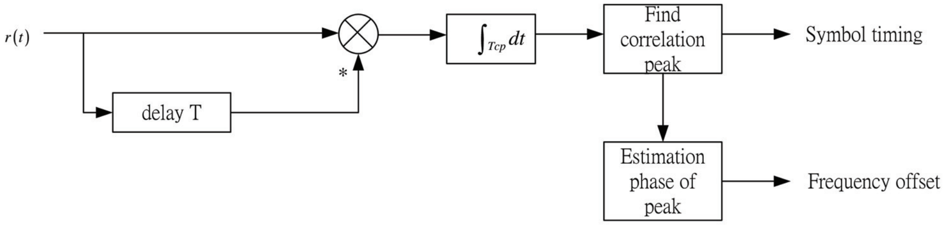


Fig. 4. Correlator using the cyclic prefix

3.2 Data-aided method

Although data-aided algorithms are not efficient for transmission, they have high estimation accuracy and a simple architecture which are especially important for packet transmission. The synchronization time needs to be as short as possible, and the accuracy must be as high as possible for high rate packet transmission (Nee & Prasad, 2000). Special OFDM training sequences in which the data is known to the receiver were developed to satisfy the requirement for packet transmission. The absolute received training signal can be exploited for synchronization, whereas non-data-aided algorithms that take advantage of cyclic extension only use a fraction signal of each symbol. In training sequence methods, the matched filter is used to estimate the symbol timing and carrier frequency offset. Fig. 5 shows a block diagram of a matched filter. The input signal is the known OFDM training sequence. The sampling interval is denoted as T . The elements of $\{c_0 \quad c_1 \quad \cdots \quad c_{N-1}\}$ are the matched filter coefficients which are the complex signals of the known training sequence. The symbol timing and carrier offset can be achieved by searching for the correlation peak accumulated from matched filter outputs.

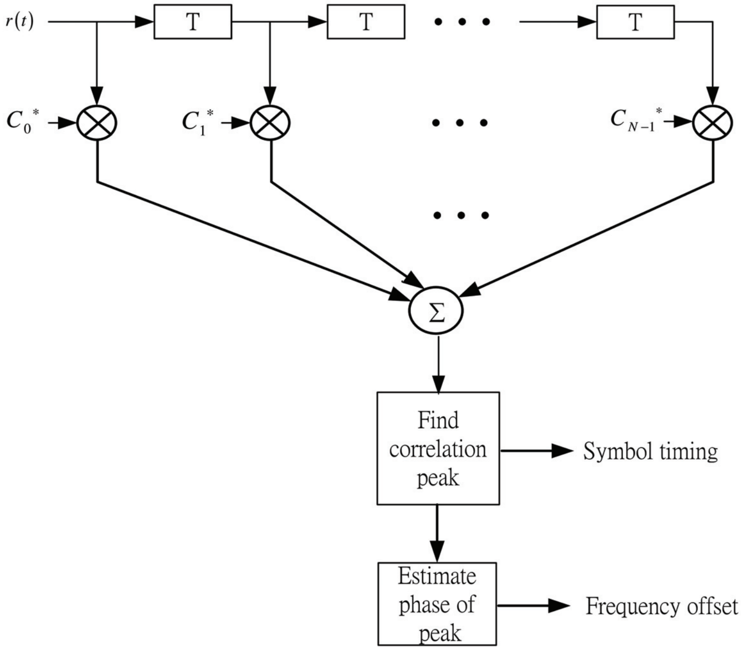


Fig. 5. Matched filter for the OFDM training sequence

4. Examples

4.1 Example 1: Non-data-aided, CP-based, fractional/fine frequency offset

According to previous researches, very high computational complexity is required for joint estimation for timing and frequency synchronization. Moreover, one estimate suffers from performance degradation caused by estimation error of the other. Thus, an effective technique is proposed (Lin, 2003).

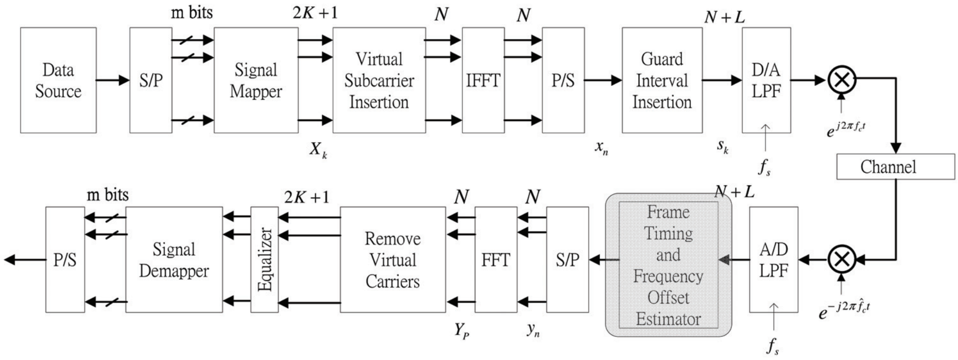


Fig. 6. The OFDM transceiver (Lin, 2003)

The proposed technique which employs a two-step method that estimates the frame timing instant and frequency offset by the maximum-likelihood (ML) estimation criterion. First, it estimates a frame timing instant such that the estimate is completely independent of the frequency offset estimation with no prior knowledge of the frequency offset; thus, a much lower estimation error of the frame timing instant is achieved by avoiding any power loss or phase ambiguity caused by frequency offset. The main reason for this arrangement is that frame timing instant estimation has to take place completely before frequency offset estimation because the latter actually requires frame timing information.

The block diagram of the OFDM system investigated here is depicted in Fig. 6. The received signal can be expressed as

$$r_k = \alpha_k s_{k-\theta} e^{j2\pi\varepsilon k/N} + n_k \tag{7}$$

where θ is the unknown delay time; α_k denotes a channel fade, which has a Rayleigh-distributed envelope and a uniformly distributed phase; ε denotes the carrier frequency offset in a subcarrier spacing; and $1/N$ is the normalized frequency. In accordance with Jake's model of a fading channel (Jakes, 1974), α_k can be expressed as a complex Gaussian random process with the autocorrelation function given as

$$E\left\{\alpha_{k_1}\alpha_{k_2}^*\right\} = J_0\left(2\pi f_D \left|k_1 - k_2\right| \frac{T_u}{N}\right) \tag{8}$$

where $E\{\cdot\}$ denotes the statistical expectation operation; $*$ denotes taking complex conjugation; $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind; f_D is the maximum Doppler frequency caused directly by relative motion; and T_u is the OFDM block duration, which actually corresponds to the time interval of an N -sample OFDM block. In a previous work (Van de Beek et al., 1997), the log-likelihood function for θ and ε can be written as

$$\begin{aligned}\Lambda(\theta,\varepsilon) &= \log f\left(\mathbf{r}\middle|\theta,\varepsilon\right)\\ &= \log\left(\prod_{k\in I} f\left(r_k, r_{k+N}\right) \prod_{k\notin I\cup I'} f\left(r_k\right)\right)\\ &= \log\left(\prod_{k\in I} \frac{f\left(r_k, r_{k+N}\right)}{f\left(r_k\right) f\left(r_{k+N}\right)} \prod_k f\left(r_k\right)\right)\end{aligned} \tag{9}$$

where $f(\cdot)$ denotes the probability density function; $\mathbf{r} = \begin{bmatrix} r_1 & r_2 & \cdots & r_{2N+L}\end{bmatrix}^T$ is the observation vector; $I = [\theta, \theta+1, \cdots, \theta+L-1]$; and $I' = [\theta+N, \theta+N+1, \cdots, \theta+N+L-1]$. It must be noted that the correlations among the samples in the observation vector are exploited to estimate the unknown parameters θ and ε, and they can be written as

$$\forall k \in I : E\left\{r_k, r_{k+m}^*\right\} = \begin{cases} E\left\{\left|r_k\right|^2\right\} = \sigma_s^2 + \sigma_n^2, & m = 0 \\ E\left\{r_k, r_{k+m}^*\right\} = \sigma_s^2 J_0\left(2\pi f_D T_u\right) e^{-j2\pi\varepsilon}, & m = N \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\sigma_s^2 = E\left[\left|s_k\right|^2\right]$ is the average signal power and $\sigma_n^2 = E\left[\left|n_k\right|^2\right]$ is the average noise power.

Because the product $\prod_k f(r_k)$ in (9) is independent of θ and ε, it can be dropped when maximizing $\Lambda(\theta,\varepsilon)$. Under the assumption that $\mathbf{r}$ is a jointly Gaussian vector and after some manipulations reported in the reference Appendix (Lin, 2003), (9) can be rewritten as

$$\begin{aligned}\Lambda(\theta,\varepsilon) &= c_1 + c_2\left[\sum_{k=\theta}^{\theta+L-1} \operatorname{Re}\left\{r_k r_{k+m}^{*} e^{-j2\pi\varepsilon}\right\} - \frac{\rho}{2}\sum_{k=\theta}^{\theta+L-1}\left(\left|r_k\right|^2 + \left|r_{k+N}\right|^2\right)\right] \\ &= c_1 + c_2\left[\operatorname{Re}\left\{\lambda_1(\theta)\right\}\cos(2\pi\varepsilon) - \operatorname{Im}\left\{\lambda_1(\theta)\right\}\sin(2\pi\varepsilon) - \rho\lambda_2(\theta)\right]\end{aligned} \tag{11}$$

where

$$\rho = \frac{E\left\{r_k r_{k+N}^{*}\right\}}{\sqrt{E\left\{\left|r_k\right|^2\right\}E\left\{\left|r_{k+N}\right|^2\right\}}} = \frac{\sigma_s^2 J_0(2\pi f_D T_u)}{\sigma_s^2 + \sigma_n^2}$$

$$c_1 = -\sum_{k=\theta}^{\theta+L-1} \log\left(1-\rho^2\right)$$

$$c_2 = \frac{2\rho}{\left(1-\rho^2\right)\left(\sigma_s^2 + \sigma_n^2\right)}$$

$$\lambda_1(\theta) = \sum_{k=\theta}^{\theta+L-1} r_k r_{k+N}^{*}$$

$$\lambda_2(\theta) = \frac{1}{2}\sum_{k=\theta}^{\theta+L-1}\left\{\left|r_k\right|^2 + \left|r_{k+N}\right|^2\right\}$$

In the above equation, it is assumed that the random frequency modulation caused by a time-varying channel fade and the phase noise of the local oscillator are negligible; thus, $\left\{r_k r_{k+N}^{*}\right\}$ has almost the same phase within the range $k \in [\theta, \theta+L-1]$; therefore, $\left\{r_k r_{k+N}^{*}\right\}$ can be coherently summed up in the term $\lambda_1(\theta)$. If the partial derivative of $\Lambda(\theta,\varepsilon)$ is taken with respect to ε, one can obtain the following equation:

$$\frac{\partial}{\partial\varepsilon}\Lambda(\theta,\varepsilon) = -2\pi c_2\left[\operatorname{Re}\left\{\lambda_1(\theta)\right\}\sin(2\pi\varepsilon) + \operatorname{Im}\left\{\lambda_1(\theta)\right\}\cos(2\pi\varepsilon)\right] \tag{12}$$

To obtain the value of $\hat{\varepsilon}$ that maximizes $\Lambda(\theta,\varepsilon)$, the above partial derivative is set to zero and equality stands only when

$$\frac{\operatorname{Re}\left\{\lambda_1(\theta)\right\}}{\cos(2\pi\varepsilon)} = \frac{\operatorname{Im}\left\{\lambda_1(\theta)\right\}}{-\sin(2\pi\varepsilon)} = \frac{1}{c_3} \tag{13}$$

where c_3 is set as a constant $1/L$ for simplicity. As a result, the carrier frequency offset estimate can be expressed as

$$\hat{\varepsilon} = -\frac{1}{2\pi}\tan^{-1}\left(\frac{\operatorname{Im}\left\{\lambda_1(\theta)\right\}}{\operatorname{Re}\left\{\lambda_1(\theta)\right\}}\right) \tag{14}$$

The carrier frequency offset estimator derived above actually requires accurate frame timing information to effectively resolve the carrier frequency offset by taking advantage of a complete cyclic prefix. As a result, accurate frame timing estimation has to be performed before a carrier frequency offset is estimated.

To develop a frame timing estimation scheme without prior knowledge of frequency offset, the log-likelihood function in (11) can be approximated as follows:

$$\begin{aligned}\Lambda(\theta,\varepsilon) &\approx c_1 + c_2\left[c_3\operatorname{Re}\{\lambda_1(\theta)\}\cdot\operatorname{Re}\{\lambda_1(\theta)\} + c_3\operatorname{Im}\{\lambda_1(\theta)\}\cdot\operatorname{Im}\{\lambda_1(\theta)\} - \rho\lambda_2(\theta)\right]\\ &= c_1 + c_2\left[c_3\left(\operatorname{Re}^2\{\lambda_1(\theta)\} + \operatorname{Im}^2\{\lambda_1(\theta)\}\right) - \rho\lambda_2(\theta)\right]\\ &= c_1 + c_2\left[c_3\left|\lambda_1(\theta)\right|^2 - \rho\lambda_2(\theta)\right]\end{aligned} \tag{15}$$

Thus, one can obtain a frame timing estimator independent of frequency offset estimation. The proposed technique provides a more practical estimate of the frame timing instant because frame timing estimation is very often performed before frequency offset is estimated or dealt with. As a result, the proposed estimator of the frame timing instant and frequency offset can be expressed as

$$\begin{cases}\text{Step 1: } \hat{\theta}_p = \arg\max\limits_{\theta}\left\{c_3\left|\lambda_1(\theta)\right|^2 - \rho\lambda_2(\theta)\right\}\\ \text{Step 2: } \hat{\varepsilon}_p = -\dfrac{1}{2\pi}\tan^{-1}\left(\dfrac{\operatorname{Im}\left\{\lambda_1\left(\hat{\theta}_p\right)\right\}}{\operatorname{Re}\left\{\lambda_1\left(\hat{\theta}_p\right)\right\}}\right)\end{cases} \tag{16}$$

Its structure is depicted in detail in Fig. 7. The proposed frame timing estimator inherently exploits the highest signal level by disregarding any phase ambiguity caused by residual error in frequency offset estimation. Therefore, the proposed technique performs frame timing estimation in a manner independent of frequency offset estimation; then, frequency offset estimation can be properly achieved in the next step by effectively taking advantage of accurate timing information.

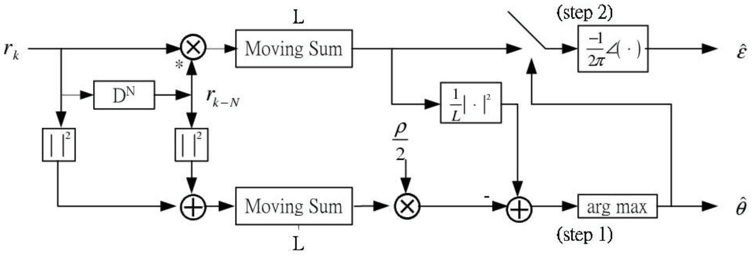


Fig. 7. The estimator (Lin, 2003)

Because the effect of fast channel fading is considered here, the proposed technique has to account for a maximum Doppler frequency f_D on the same order of $1/T_u$. Therefore, the proposed estimator of the frame timing instant is often dominated by its first term because the correlation coefficient term ρ in (16) approaches zero in such an environment. As a result, estimating of the frame timing instant can be simplified as follows to reduce the hardware complexity:

$$\hat{\theta}'_p = \arg\max_{\theta}\left\{\left|\lambda_1(\theta)\right|^2\right\} \tag{17}$$

In addition, several techniques for combining multiple frames have also been investigated (Lin, 2003) to increase the robustness of the proposed technique under low SNR conditions. Other simulation experiments show that the proposed techniques can effectively achieve lower estimation errors in frame timing and frequency offset estimation.

4.2 Example 2: Data-aided, preamble, integral/coarse frequency offset

Previous works often employ signal-estimation techniques on a time-indexed basis in the time direction. However, very few previous works have dealt with frequency-offset problems by applying a detection technique on a subcarrier-indexed basis in the frequency direction. An effective technique for frequency acquisition based on maximum-likelihood detection for mobile OFDM is proposed. The proposed technique employs a frequency-acquisition stage and a tracking stage. We mainly focus on frequency acquisition because tracking has been investigated (Lin, 2004, 2006b, 2007). By exploiting differential coherent detection of a single synchronization sequence, where a pseudonoise (PN) sequence is used as a synchronization sequence, we can prove that data-aided frequency acquisition with frequency-directional PN matched filters (MFs) reduces the probabilities of false alarm and miss on a channel with a sufficiently wide coherence bandwidth. Strict statistical analyses have been performed to verify the improvements achieved. Furthermore, the proposed technique can operate well over a channel with severe frequency-selective fading by exploiting subcarrier-level differential operation and subsequent coherent PN cross-correlation.

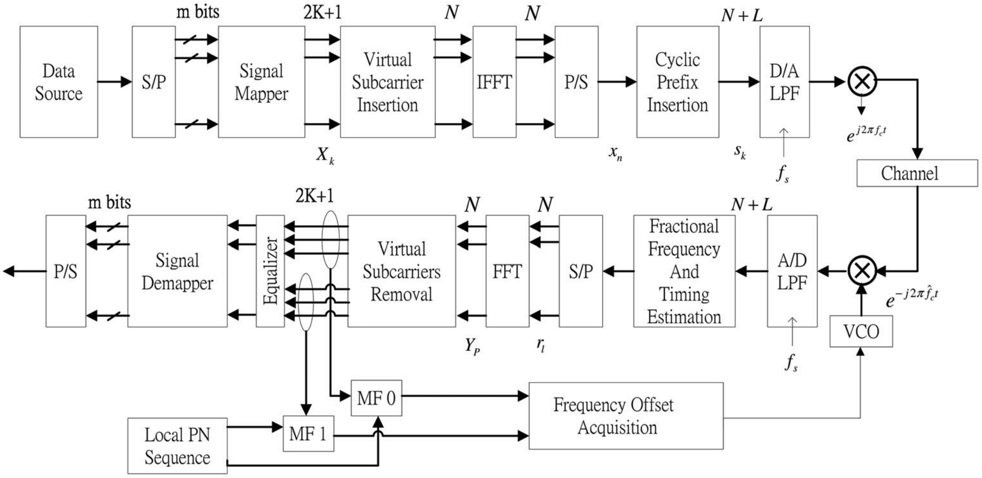


Fig. 8. The OFDM transceiver (Lin, 2006a)

In the investigated OFDM system, a PN sequence with a period N_p (say, $N_p < K$) is successively arranged to form an OFDM preamble block. The complex representation of the received baseband-equivalent signal can, thus, be written as

$$r_l = \frac{1}{\sqrt{N}} \sum_{k=-K}^{K} c_{|k|_{N_p}} \exp\left(j2\pi \frac{kl}{N}\right) \exp\left(j2\pi(d+\varepsilon)\frac{l}{N}\right) + n_l''', \quad l = 0,1,\ldots,N-1 \tag{18}$$

where l denotes the time index, the term $\exp\left(j2\pi(d+\varepsilon)(1/N)\right)$ represents the effect of the CFO that is mainly caused by instability or mismatch that occurs with the local oscillator at the front-end down-conversion process, d and ε are the integral and fractional parts of the CFO, respectively, which are normalized by the subcarrier spacing (i.e., frequency separation between any two adjacent subcarriers), $c_{|k|_{N_p}}$ is the $|k|_{N_p}$ th chip value of the PN code transmitted via the kth subchannel, whose normalized subcarrier frequency is (k/N), $|k|_{N_p}$ denotes the k modulus N_p, and n_l''' is complex white Gaussian noise. With the FFT demodulation, the pth subchannel output can be expressed as

$$\begin{aligned} Y_p &= \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} \exp\left(-j2\pi \frac{pl}{N}\right) \cdot r_l \\ &= \sum_{k=-K}^{K} c_{|k|_{N_p}} g(k+d-p+\varepsilon) \cdot \exp\left(j\pi \frac{N-1}{N}(k+d-p+\varepsilon)\right) + n_p'', \quad p = -N/2,\ldots,N/2-1 \end{aligned} \tag{19}$$

where

$$g(\upsilon) = \frac{\sin(\pi\upsilon)}{N\sin(\pi\upsilon/N)}$$

and n_p'' has a noise term. If the demodulation outputs $\left\{Y_p,\ p = 0,1,\ldots,N_p-1;\ N_p < K\right\}$ are cross-correlated with a locally generated PN sequence with a phase delay $\hat{d}$ using PN MF, then the output of the PN MF can be obtained.

$$Z_0 = \frac{\sqrt{N_p}}{\sigma_0} g\left(d-\hat{d}+\varepsilon\right) \exp\left(j\pi \frac{N-1}{N}\left(d-\hat{d}+\varepsilon\right)\right) + n_0 \tag{20}$$

The detailed derivation has been shown elsewhere (Lin, 2006a). As a result, coarse frequency offset can be detected through subcarrier acquisition. The detection procedure is equivalent to testing the following two hypotheses:

$$\begin{cases} f_{\|Z_o\|^2}\left(\eta \mid A_1, H_1\right) \sim \chi^2 \\ A_1 = g_1(\varepsilon) = \dfrac{\sin(\pi\varepsilon)}{N\sin(\pi\varepsilon/N)} = g(\varepsilon), \quad H_1 : d = \hat{d}, \\ f_{\|Z_o\|^2}\left(\eta \mid A_0, H_0\right) \sim \chi^2 \\ A_0 = g_0(\varepsilon) = g(d'+\varepsilon) = \dfrac{\sin\left(\pi(d'+\varepsilon)\right)}{N\sin\left(\pi(d'+\varepsilon)/N\right)}\Big|_{d'=d-\hat{d}\neq 0}, \quad H_0 : d \neq \hat{d} \end{cases} \tag{21}$$

where H_1 and H_0 denote the two hypothesis that the local PN sequence has been aligned (i.e., $d = \hat{d}$) and has not been aligned in-phase (i.e., $d \neq \hat{d}$), respectively, with respect to the post-FFT-demodulation PN sequence.

The previous derivations show that the major difficulty with the ordinary likelihood functions results from the very complicated probability density functions of the derived random variable, $A_0 = g_0(\varepsilon)$ and $A_1 = g_1(\varepsilon)$. Therefore, the two derived random variables A_0 and A_1 are first set to be constant for the worst cases, and thus, the (fixed) noncentrality parameters can be exploited in the likelihood functions to simplify the detection procedure. The probabilities of false alarm and miss for noncoherent detection can be written as

$$\begin{aligned} P_{fa}^{nc}(t_{nc}) &= P(S > t_{nc} | H_0) \\ &\le \int_{t_{nc}}^{\infty} f_S\left(s \middle| H_0, \max_{\varepsilon} g_0(\varepsilon)\right) ds \\ &= Q_1\left(\sqrt{\lambda_{nc,0}}, t_{nc}\right) \end{aligned} \tag{22}$$

$$\begin{aligned} P_{ms}^{nc}(t_{nc}) &= P(S \le t_{nc} | H_1) \\ &\le 1 - \int_{t_{nc}}^{\infty} f_S\left(s \middle| H_1, \min_{\varepsilon} g_1(\varepsilon)\right) ds \\ &= 1 - Q_1\left(\sqrt{\lambda_{nc,1}}, t_{nc}\right) \end{aligned} \tag{23}$$

where

$$\lambda_{nc,0} = \max_{\substack{|\varepsilon| \le 0.5 \\ d \ne \hat{d}}} \left| \frac{\sin\left(\pi\left(d - \hat{d} + \varepsilon\right)\right)}{N \sin\left(\pi\left(d - \hat{d} + \varepsilon\right)/N\right)} \frac{\sqrt{N_p}}{\sigma_0} \right|^2 \doteq 2g^2(1.5) N_p \cdot \overline{SNR}$$

$$\lambda_{nc,1} = \max_{\substack{|\varepsilon| \le 0.5 \\ d = \hat{d}}} \left| g\left(d - \hat{d} + \varepsilon\right) \frac{\sqrt{N_p}}{\sigma_0} \right|^2 = 2g^2(0.5) N_p \cdot \overline{SNR}$$

and

$$Q_{\mu/2}(a,b) = \int_b^{\infty} \frac{1}{2} \left(\frac{x}{a^2}\right)^{(\mu-2)/4} \cdot \exp\left(-\frac{x + a^2}{2}\right) I_{\frac{\mu}{2}-1}\left(\sqrt{a^2 x}\right) dx$$

is the generalized Marcum Q-function, which is defined as the complementary cumulative density function of a noncentral χ^2 random variable with μ degrees of freedom and noncentrality parameter a^2, and where t_{nc} is a design parameter representing the decision threshold of the derived noncoherent detection.

The above noncoherent detector can be further improved by a differentially coherent detection technique that consists of coherent accumulation of cross-correlations subchannel-by-subchannel by means of PN MFs. The detailed derivation has been provided elsewhere (Lin, 2006a). As a result, the probability of false alarm and miss for the proposed differentially coherent subcarrier-acquisition technique is given by

$$P_{fa}^{dc} = P(\gamma_a - \gamma_b > t_{dc} | H_0) = \int_0^{\infty} f_{\gamma_b}(s | H_0) \int_{s+t_{dc}}^{\infty} f_{\gamma_a}(\eta | H_0) d\eta ds = \int_0^{\infty} \frac{1}{2} e^{-s/2} Q_1\left(\sqrt{\lambda_{dc,o}}, s + t_{dc}\right) ds \tag{24}$$

$$P_{fa}^{dc} = P(\gamma_a - \gamma_b \le t_{dc} | H_1) = 1 - \int_0^{\infty} \frac{1}{2} e^{-s/2} Q_1\left(\sqrt{\lambda_{dc,1}}, s + t_{dc}\right) ds \tag{25}$$

where

$$\lambda_{dc,0} = \Lambda^2 \Big|_{\substack{H_0 \\ d \neq \hat{d}}} = 4g^2(1.5)N_p\overline{SNR}$$

$$\lambda_{dc,1} = \Lambda^2 \Big|_{\substack{H_1 \\ d = \hat{d}}} = 4g^2(0.5)N_p\overline{SNR}$$

and t_{dc} is a design parameter denoting the decision threshold when the above differentially coherent detection is used.

It can be easily seen from simulation results (Lin, 2006a) that no matter what values of the decision threshold are chosen, the proposed techniques can achieve sufficiently low probabilities of false alarm and miss and that differentially coherent detection can achieve lower probabilities than its noncoherent counterpart. The main reason for this difference is that differentially coherent detection primarily tests two more distantly separated distributions than does the noncoherent detection.

Although the previous derivations were conducted only on an AWGN channel, similar results and conclusions hold for a flat-fading channel or in an environment whose coherence bandwidth is wide enough to accommodate several subchannels. The relative contexts are shown completely in the reference paper (Lin, 2006a).

5. Synchronization in LTE/LTE-A systems

5.1 Introduction

Requirement of transmission data rate grows up rapidly as time flies. The Long Term Evolution (LTE) specification proposed by 3rd Generation Partnership Project (3GPP) has a significant influence on recent wireless communications. LTE communication systems are expected to achieve a data rate of 100 Mb/s on downlink and 50 Mb/s on uplink transmissions; it can also provide flexible bandwidths of 1.4, 2.5, 5, 10, 15 and 20 MHz. An LTE communication is based on the OFDM techniques and adopts single-carrier frequency-division multiple access (SC-FDMA) on uplink transmission and OFDMA on downlink transmission. It is clear that LTE can provide a high data rate, robust performance over multipath fading channels and high spectral efficiency. However, an LTE system has a major drawback: it is sensitive to frequency error as OFDM systems. Timing and frequency synchronization is a key component for initial synchronization of an LTE system. For a link initiative, a mobile station has to search for a base-station by means of synchronization sequences, which are broadcasted in all directions in which the station provides coverage. This search is called *cell search* in cellular systems. In the cell search, a sector search must be performed at first. The following tasks comprise the sector search: coarse timing alignment, fine timing synchronization, fine frequency offset compensation, coarse frequency offset detection, and sector identification.

5.2 LTE frame structure

An LTE supports 504 different cell identifications. To avoid the high complexity of a cell search procedure, these cell identifications are categorized into 168 cell-identification groups, $N_{ID}^{(1)}$; additionally, each cell-identification group contains three identities, $N_{ID}^{(2)}$. Therefore, cell identification (ID) can be stated as $N_{ID}^{cell} = 3N_{ID}^{(1)} + N_{ID}^{(2)}$. Initially, the sector of the received signal has to be identified. Then, the cell that can provide service must be identified. After the above procedure is completed, communication can begin. An LTE supports two synchronization signals for the cell search procedure. One is the primary synchronization signal (P-SCH), and the other is the secondary synchronization signal (S-SCH). P-SCH and S-SCH are inserted into the last two OFDM symbols in the first slot of the sub-frame zero and sub-frame five, where the frame structure is shown in Fig. 9. The P-SCH signal is transmitted twice in each 10-ms frame. It can provide frame timing synchronization with a tolerance of 5 ms. The main goal of the P-SCH is to conduct timing synchronization, coarse frequency-offset detection and sector identification. Each frame has a pair of S-SCH signals that can be chosen from the 168 different cell identifications. Therefore, the S-SCH signal is used to determine the cell ID.

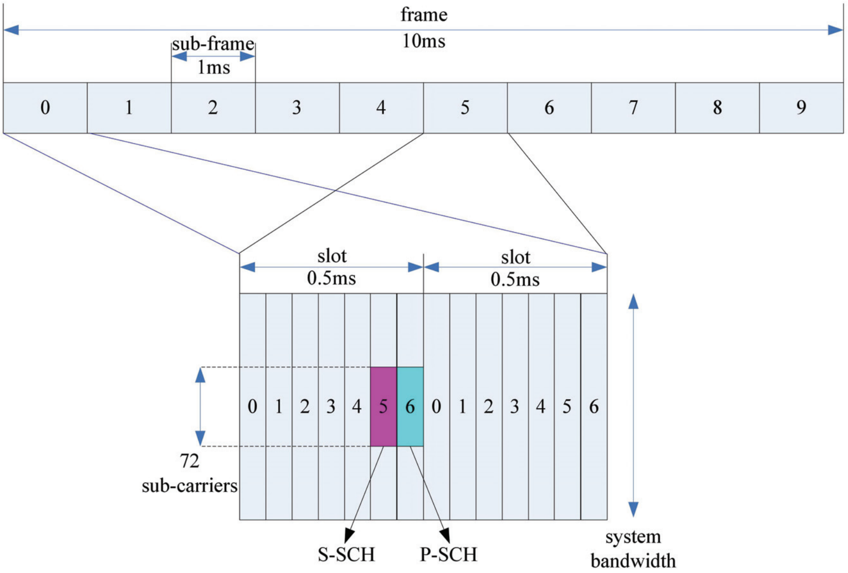


Fig. 9. LTE frame structure

The frame structure of the LTE system is depicted in Fig. 9, and the length of each frame is 10 ms. Each frame is divided into ten 1-ms sub-frames. Each sub-frame contains two slots with lengths 0.5 ms. Additionally, each slot consists of seven symbols, and each symbol contains 2048 samples. The zeroth and fifth sub-frame convey P-SCH and S-SCH signals. According to the LTE specification, the CP length is 160 samples in the first symbol of a slot and 144 samples in the other 6 symbols of the slot. When the occupied bandwidth is 20 MHz, the system parameters are as follows: the sampling rate is 30.72 MHz, the FFT size is 2048, and the subcarrier spacing is 15 KHz. The synchronization signals occupy only the central 72 sub-carriers of the 2048 sub-carriers. Both boundaries of the band conveying the synchronization signals accommodate 5 null subcarriers. Therefore, the synchronization signals only adopt 62 subcarriers.

5.3 P-SCH signal

The number of physical-layer cell identifications is very large. To avoid high complexity in the cell search, the cell identifications are partitioned into three sets, physical-layer cell-identification group $N_{ID}^{(2)}$ or sector. The P-SCH signal is composed of three Zadoff-Chu (ZC) sequences with lengths of 62 in the frequency domain. Each sequence represents a sector identification. The ZC sequences employed in the LTE (3GPP LTE, 2005) are written as

$$d_u(n) = \begin{cases} e^{-j\frac{\pi un(n+1)}{63}}, & n = 0,1,....,30 \\ e^{-j\frac{\pi u(n+1)(n+2)}{63}}, & n = 31,32,....,61. \end{cases} \tag{26}$$

$N_{ID}^{(2)}$	Root index u
0	25
1	29
2	34

Table 1. Root index u of sector identification (3GPP LTE, 2005)

where u is the root index for which values are set to 25, 29, and 34, which correspond to $N_{ID}^{(2)}$ =0,1 or 2, respectively. A ZC sequence is a chirp-like sequence and is symmetric both in the time domain and frequency domain. The sequence has good correlation properties. Therefore, the P-SCH signal employing the ZC sequence is utilized to help coarse timing synchronization and frequency-offset detection.

5.4 Cell search method

Research regarding sector search in LTE systems has been studied extensively (Chen et al., 2009, Manolakis et al., 2009, Tsai et al., 2007). Three methods were studied previously (Tsai et al., 2007). They mainly take advantage of auto-correlation, cross-correlation and hybrid detection. The first method adopts auto-correlation to search for P-SCH location by taking advantage of the repetitions of P-SCH sequences. Coarse frequency-offset acquisition depends on the output of the auto-correlator. Its main advantage is low complexity, but the timing detection is inevitably distorted on signals with low SNR. The second method employs cross-correlation between the received signal and the locally-generated P-SCH to detect timing and frequency offset. Additionally, the cross-correlation can be divided into several segments to counter any effect caused by a high frequency offset. The method has a trustworthy timing detection, but its complexity is higher than auto-correlation detection. Hybrid detection combines advantages of auto-correlation and cross-correlation. Its complexity is less than that employing cross-correlation detection. The auto-correlation detection obtains coarse timing and frequency offset, and compensates for the frequency error. Then, the accurate timing can be obtained by exploiting cross-correlation.A previous study (Manolakis et al., 2009) used maximum likelihood (ML) criterion to estimate the fractional frequency offset and the OFDM symbol timing; its performance is improves by averaging 8 OFDM symbols. Next, cross-correlation between the three P-SCH sequences and the received signal is obtained; and the sector ID can be determined by selecting the highest cross-correlation according to the orthogonality among the used Zadoff-Chu sequences.

5.5 Carrier aggregation

Carrier aggregation is one of the most important technologies in the new LTE-Advanced standards. This technique will also play a significant role for 4G communication systems. By using carrier aggregation, a peak data rate up to 1 Gb/s is possible in future 4G mobile communications. Because of the flexibility of effective transmission, the user can exploit numerous carriers at the same time. In addition, these carriers may lie in the same or different band and may have different bandwidths. Carrier aggregation provides diverse combinations and flexible spectrum usability and has attracted attention. Carrier aggregation techniques can be classified into two categories: continuous and discontinuous as shown in Fig. 10. These two categories can be subdivided into three types: intraband contiguous, intraband discontinuous and interband. A diagram describes their difference in Fig. 11.

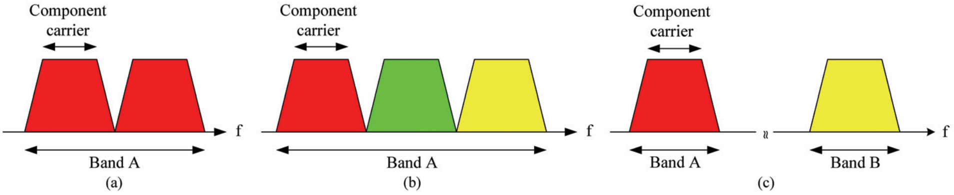


Fig. 10. Carrier aggregation types: (a) intraband contiguous; (b) intraband discontinuous; (c) interband (Iwamura et al., 2010)

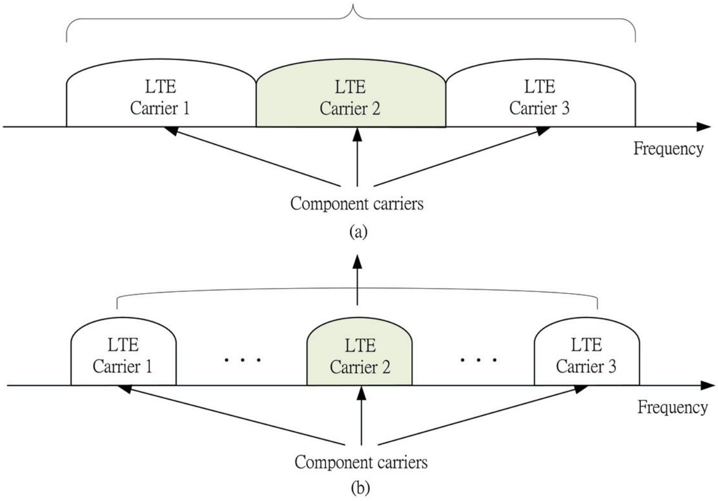


Fig. 11. Carrier aggregation categories: (a) continuous; (b) discontinuous (Yuan et al., 2010)

6. Summary

In this chapter, the authors intend to introduce the OFDM communication systems and take care of the main issue, frequency offset, can lead to severe performance degradation. Two classifications of synchronization techniques are introduced. Several novel techniques have been thoroughly discussed in great detail in this chapter. LTE/LTE-A systems have been chosen as candidates for 4G mobile communication. The concept of LTE-LTE-A systems is mentioned in the end of this chapter.

7. References

3GPP LTE (2005). TS 36.211 V8.3.0: Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 8)

Bolcskei, H. (2001). Blind estimation of symbol timing and carrier frequency offset in wireless OFDM systems, *IEEE Transactions on Communications*, Vol.49, No.6, (June 2001), pp.988-999

Chen, Y., Wen, X., Zheng, W. & Lin, X. (2009). Symbol timing estimation and sector detection algorithm based on LTE TDD system, *Proceedings of IEEE Network Infrastructure and Digital Content Conference, 2009 (IC-NIDC 2009)*, Beijing, China, pp.828-832.

Classen, F. & Meyr, H. (1994). Frequency synchronization algorithms for OFDM systems suitable for communication over frequency selective fading channels, *Proceedings of IEEE Vehicular Technology Conference, 1994 (VTC'94)*, Stockholm, Sweden, pp. 1655-1659

Daffara, F. & Chouly, A. (1993). Maximum likelihood frequency detectors for orthogonal multicarrier systems, *Proceedings of IEEE Communications Conference, 1993 (ICC'93)*, Geneva, Switzerland, pp. 766-771

Daffara, F. & Adami, O. (1995). A new frequency detector for orthogonal multicarrier transmission techniques, *Proceedings of IEEE Vehicular Technology Conference, 1995 (VTC'95)*, Chicago, USA, pp. 804-809

Dahlman, E., Parkvall, S., Skold, J. & Beming, P. (2007) *3G Evolution HSPA and LTE for Mobile Broadband,* Academic Press

Iwamura, M., Etemad, K., Fong M.-H., Nory, R. & Love, R. (2010) Carrier aggregation framework in 3GPP LTE-advanced [WiMAX/LTE update], *IEEE Communications Magazine*, Vol.48, No.8, (August 2010), pp.66-67

Jakes, W. C. & Cox, D. C. (1994). *Microwave Mobile Communications*. Wiley-IEEE Press

Kapoor, S., Marchok, D. J. & Huang, Y.-F. (1998). Pilot assisted synchronization for wireless OFDM systems over fast time varying fading channels, *Proceedings of IEEE Vehicular Technology Conference, 1998 (VTC'98)*, Ottawa, Canada, pp. 2077-2080

Lin, J.-C. (2002a). Noncoherent sequential PN code acquisition using sliding correlation for chip-asynchronous DS/SS, *IEEE Transactions on Communications*, Vol.50, No.4, (April 2002), pp.664-676

Lin, J.-C. (2002b). Differentially coherent PN code acquisition with full-period correlation in chip-asynchronous DS/SS receivers, *IEEE Transactions on Communications*, Vol.50, No.5, (May 2002), pp.698-702

Lin, J.-C. (2002c). Differentially coherent PN code acquisition based on a matched filter for chip-asynchronous DS/SS communications, *IEEE Transactions on Vehicular Technology*, Vol,51, No.6, (November 2002), pp.1596-1599

Lin, J.-C. (2003). Maximum-likelihood frame timing instant and frequency offset estimation for OFDM communication over a fast Rayleigh-fading channel, *IEEE Transactions on Vehicular Technology*, Vol,52, No.4, (July 2003), pp.1049-1062

Lin, J.-C. (2004). Frequency offset estimation by differentially coherent frequency error characterization for OFDM wireless communications, *Proceedings of IEEE Communications Conference, 2004 (ICC'04)*, Paris, France, pp. 2387-2391.

Lin, J.-C. (2005). Frequency offset acquisition based on subcarrier differential detection for OFDM communications on doubly-selective fading channel, *Proceedings of IEEE Communications Conference, 2005 (ICC'05)*, Seoul, Korea, pp. 1952-1956.

Lin, J.-C. (2006a). Coarse frequency-offset acquisition via subcarrier differential detection for OFDM communications, *IEEE Transactions on Communications*, Vol.54, No.8, (August 2006), pp.1415-1426

Lin, J.-C. (2006b). Frequency offset estimation technique based on error characterization for OFDM communications on time-varying multipath fading channels, *Proceedings of IEEE Communications Conference, 2006 (ICC'06)*, Istanbul, Turkey, pp.2911-2916.

Lin, J.-C. (2007). A frequency offset estimation technique based on frequency error characterization for OFDM communications on multipath fading channels, *IEEE Transactions on Vehicular Technology*, Vol,56, No.3, (May 2007), pp.1209-1222

Luise, M. & Reggiannini, R. (1996). Carrier frequency acquisition and tracking for OFDM systems, *IEEE Transactions on Communications*, Vol.44, No.11, (November 1996), pp.1590-1598

Lv, T., Li, H. & Chen, J. (2005) Joint estimation of symbol timing and carrier frequency offset of OFDM signals over fast time-varying multipath channels, *IEEE Transaction on Signal Processing*, Vol.53, No.12, (December 2005), pp.4526-4535

Manolakis, K., Gutierrez Estevez, D. M., Jungnickel, J., Xu, W. & Drewes, C. (2009) A closed concept for synchronization and cell Search in 3GPP LTE systems, *Proceedings of IEEE Wireless Communications and Networking Conference, 2009 (WCNC 2009)*, Budapest, Hungary, pp. 1-6.

Moose, P. H. (1994). A technique for orthogonal frequency division multiplexing frequency offset correction, *IEEE Transactions on Communications*, Vol.42, No.10, (October 1994), pp.2908-2914

Nee, V. R. & Prasad, R., (2000). *OFDM for Wireless Multimedia Communications*, Artench House

Okada, M., Hara, S., Komaki, S. & Morinaga, N. (1996). Optimum synchronization of orthogonal multi-carrier modulated signals, *Proceedings of IEEE Personal, Indoor and Mobile Radio Communications Conference, 1996 (PIMRC'96)*, Taipei, Taiwan, pp. 863-867

Park, B., Ko, E., Cheon, H., Kang, C. & Hong, D. (2001). A Blind OFDM synchronization algorithm based on cyclic correlation, *IEEE Signal Processing Letters*, Vol.11, No.2, (February 2004), pp.83-85.

Popovic, B. M. (1992). Generalized chirp-like polyphase sequences with optimum correlation properties, *IEEE Transactions on Information Theory*, Vol.38, No.4, (July 1992), pp.1406-1409

Van de Beek, J.-J., Sandell, M. & Borjesson, P. O. (1997). ML estimation of time and frequency offset in OFDM systems, *IEEE Transaction on Signal Processing*, Vol.45, No.7, (July 1997), pp. 1800-1805

Tsai, Y., Zhang, G., Grieco, D. & Ozluturk, F. (2007). Cell searrch in 3GPP Long Term Evolution systems, *IEEE Vehicular Technology Magazine*, Vol.2, No.2, (June 2007), pp.23-29

Warner, W. D. & Leung, C. (1993). OFDM/FM frame synchronization for mobile radio data communications, *IEEE Transactions on Vehicular Technology*, Vol,42, No.3, (August 1993), pp.302-313

Yuan, G., Zhang, X., Wang, W. & Yang, Y. (2010). Carrier Aggregation for LTE-advanced mobile communication systems, *IEEE Transaction on Communication Magazine*, Vol.48, No.2, (February 2010), pp.88-93

3

ICI Reduction Methods in OFDM Systems

Nadieh M. Moghaddam and Mohammad Mohebbi
Iran University of Science and Technology
Iran

1. Introduction

The principles of multicarrier modulation have been in existence for several decades. However, in recent years these techniques have quickly moved out of textbooks and into practice in modern communications systems in the form of orthogonal frequency division multiplexing (OFDM). OFDM is a special form of multicarrier modulation technique which is used to generate waveforms that are mutually orthogonal and then distributes the data over a large number of carriers that are spaced apart at precise frequencies. This spacing provides the "orthogonality" in this technique which prevents the demodulators from seeing frequencies other than their own. In an OFDM scheme, a large number of orthogonal, overlapping, narrow band subcarriers are transmitted in parallel. These carriers divide the available transmission bandwidth. The separation of the subcarriers is such that there is a very compact spectral utilization. With OFDM, it is possible to have overlapping sub channels in the frequency domain (Figure 1), thus increasing the transmission rate.

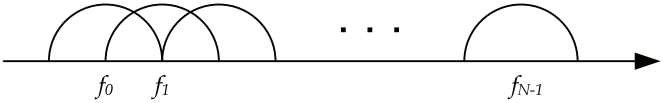


Fig. 1. Power spectrum of the transmitted signal

In order to avoid a large number of modulators and filters at the transmitter and complementary filters and demodulators at the receiver, it is desirable to be able to use modern digital signal processing techniques, such as fast Fourier transform (FFT).
OFDM is a promising candidate for achieving high data rates in mobile environment because of its multicarrier modulation technique and ability to convert a frequency selective fading channel into several nearly flat fading channels.
This technology has been chosen as the transmission method of many standards, such as Digital Subscribe Line (DSL), European Digital Audio and Video Broadcasting terrestrial (DAB/DVB-T), European HIPERLAN/2 and IEEE 802.11 a/g for wireless local area networks (WLAN), Worldwide Interoperability for Microwave Access (WiMAX), etc. However, OFDM systems exhibit a sensitivity to phase noise higher than single carrier modulations due to its long symbol period. Because carriers are kept very close to each other, OFDM is very sensitive to distortion that may remove the orthogonality between carriers. The crystal oscillator used in a mixer generates phase noise. It can also be caused by

AWGN present at the input of a Phase Locked Loop (PLL) in a coherent receiver. Phase noise can cause several types of signal degradation that are usually very difficult to quantify analytically. When the modulation experiences phase noise, it encounters two problems: 1) a common phase rotation over all the carrier frequencies which rotate the entire signal space for a given OFDM symbol and 2) inter-carrier interference due to the loss of orthogonality between subcarriers. Especially, the ICI seriously degrades system predominance because it may break down the orthogonality between subcarriers.

There have been many previous works on the phase noise, frequency offset and reduction of ICI. Among them the following methods are discussed and compared in this chapter. In the next section the OFDM system is introduced and its benefits along with its drawbacks are analyzed. ICI reduction methods such as pulse shaping and self-cancellation are given in section 3 and the last section concludes the chapter.

2. OFDM system

Figure 2 shows the block diagram of a typical OFDM system. The transmitter section converts digital data to be transmitted, into a mapping of subcarrier amplitude and phase. It then transforms this spectral representation of the data into the time domain using an Inverse Discrete Fourier Transform (IDFT). The Inverse Fast Fourier Transform (IFFT) performs the 20 same operations as an IDFT, except that it is much more computationally efficient, and so is used in all practical systems. In order to transmit the OFDM signal the calculated time domain signal is then mixed up to the required frequency. The receiver performs the reverse operation of the transmitter, mixing the RF signal to base band for processing, then using a Fast Fourier Transform (FFT) to analyze the signal in the frequency domain. The amplitude and phase of the subcarriers is then picked out and converted back to digital data. The IFFT and the FFT are complementary function and the most appropriate term depends on whether the signal is being received or generated. In cases where the signal is independent of this distinction then the term FFT and IFFT is used interchangeably.

The high data rate serial input bit stream is fed into serial to parallel converter to get low data rate output parallel bit stream. Input bit stream is taken as binary data. The low data rate parallel bit stream is modulated in Signal Mapper. Modulation can be BPSK, QPSK, QAM, etc. The modulated data are served as input to inverse fast Fourier transform so that each subcarrier is assigned with a specific frequency. The frequencies selected are orthogonal frequencies. In this block, orthogonality in subcarriers is introduced. In IFFT, the frequency domain OFDM symbols are converted into time domain OFDM symbols. Guard interval is introduced in each OFDM symbol to eliminate inter symbol interference (ISI). All the OFDM symbols are taken as input to parallel to serial data. These OFDM symbols constitute a frame. A number of frames can be regarded as one OFDM signal. This OFDM signal is allowed to pass through digital to analog converter (DAC). In DAC the OFDM signal is fed to RF power amplifier for transmission. Then the signal is allowed to pass through additive white Gaussian noise channel (AWGN channel). At the receiver part, the received OFDM signal is fed to analog to digital converter (ADC) and is taken as input to serial to parallel converter. In these parallel OFDM symbols, Guard interval is removed and it is allowed to pass through Fast Fourier transform. Here the time domain OFDM symbols are converted into frequency domain. After this, it is fed into Signal Demapper for demodulation purpose. And finally the low data rate parallel bit stream is converted into high data rate serial bit stream which is in form of binary.

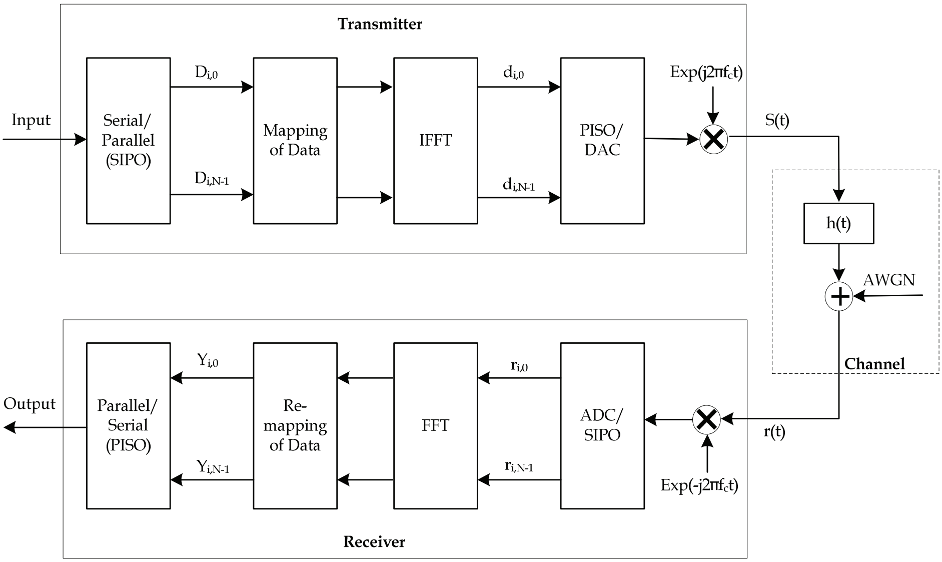


Fig. 2. OFDM system implementation

By the insertion of an extra guard interval between successive OFDM symbols the Inter Symbol Interference (ISI) can be avoided. The guard interval could be a section of all zero samples transmitted in front of each OFDM symbol and its duration should be more than the channel delay spread (L_c). It should be considered that in practical systems the guard interval is not used. Instead, Cyclic Prefix (CP) is inserted to combat the multipath-channel by making the channel estimation simple. The cyclic prefix is a replica of the last L_p samples of the OFDM symbol where $L_p > L_c$. Because of the way in which the cyclic prefix was formed, the cyclically-extended OFDM symbol now appears periodic when convolved with the channel. An important result is that the effect of the channel becomes multiplicative.
For the better understanding of this issue assume that the impulse response of the channel is $h_0, h_1, \ldots, h_{L_c}$ and the *i*-th transmitted signal block in the output of IFFT block is $d_{i,0}, d_{i,1}, \ldots, d_{i,N-1}$. In this condition the cyclic prefix would be $d_{i,N-L_c}, d_{i,N-L_c+1}, \ldots, d_{i,N-1}$. The symbols of the received baseband signal after the transmission through the channel are equal to:

$$\begin{aligned} r_{i,-L_c} &= h_0 d_{i,N-L_c} + h_1 d_{i-1,N-1} + \cdots + h_{L_c} d_{i-1,N-L_c} \\ r_{i,-L_c+1} &= h_0 d_{i,N-L_c+1} + h_1 d_{i-1,N-L_c} + \cdots + h_{L_c} d_{i-1,N-L_c+1} \\ &\vdots \\ r_{i,0} &= h_0 d_{i,0} + h_1 d_{i,N-1} + \cdots + h_{L_c} d_{i,N-L_c} \\ r_{i,1} &= h_0 d_{i,1} + h_1 d_{i,0} + \cdots + h_{L_c} d_{i,N-L_c+1} \\ &\vdots \\ r_{i,N-1} &= h_0 d_{i,N-1} + h_1 d_{i,N-2} + \cdots + h_{L_c} d_{i,N-L_c-1} \end{aligned} \tag{1}$$

At the receiver the first L_c+*1* symbols are discarded and the *N* remained symbols are demodulated using an *N*-point FFT. So the data on the *k*-th subcarrier is as follows:

$$Y_{i,k} = \sum_{n=0}^{N-1} r_{i,n} e^{-j\frac{2\pi nk}{N}} = D_{i,k} \sum_{n=0}^{Lc} h_n e^{-j\frac{2\pi nk}{N}} = D_{i,k} H_k \qquad (2)$$

where H_k is the channel impulse response in the frequency domain. It can also be considered as the channel gain on the k-th subcarrier. $D_{i,k} = \sum_{n=0}^{N-1} d_{i,n} e^{-j\frac{2\pi nk}{N}}$ is the input symbols of the IFFT block at the transmitter as:

$$d_{i,n} = N\_IFFT\{D_{i,k}\} \qquad (3)$$

It can be seen that the main symbol stream $\{D_{i,k}\}$ could be detected with the estimation of the channel coefficient at the receiver, while there is no Inter-Symbol Interference (ISI) or Inter-Carrier Interference (ICI).

Thus, a multipath channel is converted into scalar parallel sub-channels in frequency domain, thereby simplifying the receiver design considerably. The task of channel estimation is simplified, as we just need to estimate the scalar coefficients H_k for each sub-channel and once the values of $\{H_k\}$ are estimated, for the duration in which the channel does not vary significantly, merely multiplying the received demodulated symbols by the inverse of H_k yields the estimates of $\{D_{i,k}\}$.

The benefits of OFDM are high spectral efficiency, resiliency to RF interference, and lower multi-path distortion. This is useful because in a typical terrestrial broadcasting scenario there are multipath-channels (i.e. the transmitted signal arrives at the receiver using various paths of different length). Since multiple versions of the signal interfere with each other (inter symbol interference (ISI)) it becomes very hard to extract the original information. With the rapid growth of digital communication in recent years, the need for high speed data transmission is increased. Moreover, future wireless systems are expected to support a wide range of services which includes video, data and voice. OFDM is a promising candidate for achieving high data rates in mobile environment because of its multicarrier modulation technique and ability to convert a frequency selective fading channel into several nearly flat fading channels.

However, there are some non-idealities which can affect the performance of an OFDM system. These non-idealities are as follows:

Noise: Like other communication systems, the performance of an OFDM system is affected by different kind of noise such as uniform noise (AWGN), non-uniform noise (colored noise), and impulse noise.

LO phase offset: This condition occurs when there is a difference between the phase of the output LO and the phase of the received signal.

FFT window location offset: In practice, a correlation is often used with a known preamble sequence located at the beginning of the transmission. This correlation operation aids the receiver in synchronizing itself with the received OFDM symbol boundaries. However, inaccuracies still remain, and they manifest themselves as an offset in the FFT window location. The result is that the N symbols sent to the FFT will not line up exactly with the corresponding OFDM symbol.

Sampling frequency offset: A sampling frequency offset occurs when the A/D converter output is sampled either too fast or too slow.

Non-linearity in the transmitter and receiver circuits: All transmitters and receivers in communication systems contain devices such as amplifiers which are often designed to be non-linear in order to minimize power consumption. On the other hand, an OFDM signal is made up of multiple simultaneous signals that, for a given average power, have a higher peak signal level. Thus, OFDM signals result in an increase in the peak-to-average ratio (PAR) of the signal. Because of the non-linear transfer functions of amplifiers, these higher peak amplitude levels will create more severe distortion than a single carrier case even if the average power levels of each are the same.

Phase noise: OFDM systems are very sensitive to phase noise caused by oscillator instabilities in both the transmitter and the receiver. Without loss of generality, in this study the local oscillator in the receiver will be considered as the phase noise source. As mentioned before, the modulated subcarriers overlap spectrally, but since they are orthogonal over symbol duration, they can be easily recovered as long as the channel and other non-idealities do not destroy the orthogonality. An unwindowed OFDM system has rectangular symbol shapes. Therefore, in the frequency domain the individual sub-channels will have the form of sinc functions where the first sidelobe is only some 13 dB below the main lobe of the subcarrier (Figure 3). A practical oscillator has spectral components around the centre frequency. These components cause the loss of orthogonality of the OFDM carriers. In the frequency domain it can be viewed as interference caused by the high sidelobes of the adjoining carriers on a particular subcarrier.

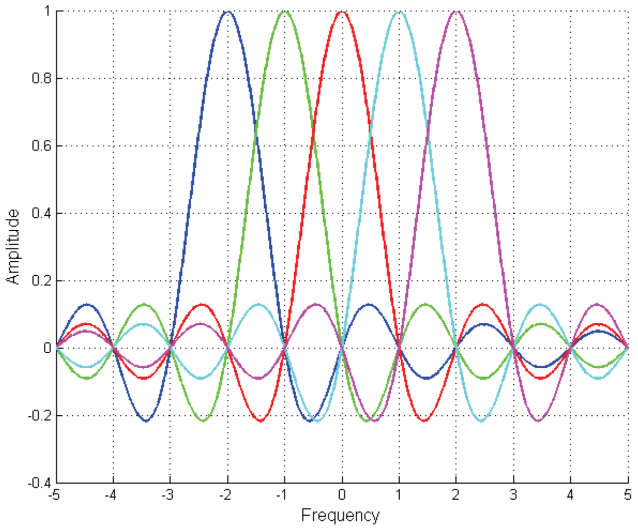


Fig. 3. OFDM spectrum with 5 subcarriers

In this study the phase noise is resolved into two components, namely the Common Phase Error (CPE), which affects all the sub-channels equally and the Inter-Carrier Interference (ICI), which is caused by the loss of orthogonality of the subcarriers.

As described above, the IFFT of $\mathrm{D_i}$ is given by $\mathrm{d_i}$. The k-th sample of d_i can be represented by:

$$d_{i,k} = \frac{1}{N}\sum_{n=0}^{N-1} D_{i,n} e^{\left(j2\pi kn/_{N}\right)} \qquad k = 0,1,\dots,N-1 \tag{4}$$

The phase noise is modeled as a phasor $\mathrm{e}^{j\theta(n)}$, where the phase noise process $\theta(n)$ is zero-mean and wide-sense stationary with a finite variance σ_θ^2. An approximation for the PSD of a free-running oscillator can be found in [Robertson & Kaiser, 1995] is as follows:

$$S_\theta(f) = 10^{-c} + \begin{cases} 10^{-a} & |f| \le f_l \\ 10^{(f_l-|f|)(\frac{b}{f_h-f_l})-a} & |f| > f_l \end{cases} \tag{5}$$

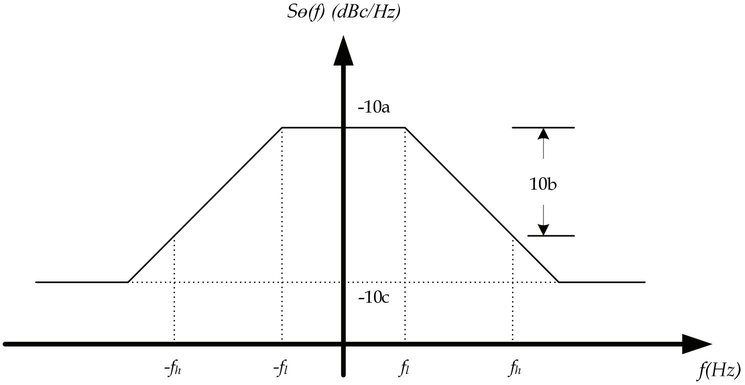


Fig. 4. Phase noise PSD of a typical oscillator

Parameter c determines the noise floor of the oscillator and a determines the noise level in the frequency ranges from the center frequency to f_l. Parameter b gives the noise fall off rate from the noise floor at f_l to the noise level at f_h. (See Figure 4).

The demodulated data symbol of the k-th subcarrier of the i-th OFDM symbol $Y_{i,k}$ with the consideration of phase noise is given by:

$$Y_{i,k} = \sum_{n=0}^{N-1} r_{i,n} e^{-j2\pi kn/N} e^{j\theta(n)} \qquad k = 0,1,\dots,N-1 \tag{6}$$

Using (2), (6) and considering AWGN,

$$Y_{i,k} = \frac{1}{N} \sum_{n=0}^{N-1} \left\{ \left\{ \sum_{l=0}^{N-1} D_{i,l} H_l e^{j2\pi ln/N} \right\} e^{j\theta(n)} \right\} e^{-j2\pi kn/N} + W_k \tag{7}$$

where W_k is the contribution due to AWGN on the k-th subcarrier. We can further simplify (7) as:

$$Y_{i,k} = \frac{1}{N} H_k D_{i,k} \sum_{n=0}^{N-1} e^{j\theta(n)} + \frac{1}{N} \sum_{\substack{l=0 \\ l\neq k}}^{N-1} H_l D_{i,l} \sum_{n=0}^{N-1} e^{j\theta(n)} e^{-j2\pi(k-l)n/N} + W_k \tag{8}$$

The first term on the right-hand side of (8) rotates the useful component $H_k D_{i,k}$ of each subcarrier by an equal amount and is independent of the particular subchannel concerned, k.

This is commonly known as the Common Phase Error (CPE). The second term is the Inter-Carrier Interference (ICI) caused by contributions from all subcarriers $l \neq k$ on k due to the loss of orthogonality. Unlike the CPE, ICI is not easy to estimate.

Local oscillator frequency offset: It drives from the difference between LO frequency at the transmitter and the receiver. In addition to phase noise, the frequency offset which is caused by the high Doppler spread and the mismatch in the oscillator frequency, produces inter-carrier interference (ICI).

For a block of data, the modulated signal at the n-th instant of time can be written as:

$$d_n = \frac{1}{N}\sum_{k=0}^{N-1} D_k\, e^{j2\pi kn/N} \quad ,n = 0,1,\dots,N-1 \tag{9}$$

The signal at the receiver, after passing through a frequency selective fading channel is expressed as:

$$r_n = \sum_{l=0}^{Lc} d_{n-l}\, h_l + w_n = \sum_{l=0}^{Lc}\frac{1}{N}\sum_{k=0}^{N-1} D_k\, e^{j2\pi kn/N}\, h_l + w_n \tag{10}$$

At the receiver, the signal is mixed with a local oscillator signal which is Δf above the correct carrier frequency. The signal Y_k received at the k-th subcarrier after performing the FFT is expressed as:

$$\begin{aligned} Y_k &= \sum_{n=0}^{N-1} r_n e^{-\frac{j2\pi kn}{N}} e^{\frac{j2\pi n\varepsilon}{N}} = D_k H_k \frac{1}{N}\sum_{n=0}^{N-1} e^{\frac{j2\pi n\varepsilon}{N}} + \sum_{\substack{l=0\\ l\neq k}}^{N-1} D_l H_l \frac{1}{N}\sum_{n=0}^{N-1} e^{\frac{j2\pi n(l+\varepsilon-k)}{N}} + w'_k \\ &= D_k H_k I(0) + \sum_{\substack{l=0\\ l\neq k}}^{N-1} D_l H_l I(l-k) + w'_k\,,\quad k = 0,1,\dots,N-1 \end{aligned} \tag{11}$$

Where $\varepsilon = \Delta f \times T_s$ is the frequency offset normalized to the OFDM symbol rate $1/T_s$. The T_s denotes the OFDM symbol duration excluding the guard interval. The sequence $I\ (l-k)$ is defined as the ICI coefficient between l-th and k-th subcarriers, which can be expressed as:

$$I(l-k) = \frac{sin\ (\pi(l+\varepsilon-k))}{Nsin(\frac{\pi}{N}(l+\varepsilon-k))} exp\ (j\pi(1-\frac{1}{N})(l+\varepsilon-k)) \tag{12}$$

The first term in the right-hand side of (11) represents the desired signal. Without frequency error ($\varepsilon\ = 0$), $I\ (0)$ takes its maximum value. The second term is the ICI components, which as ε becomes larger, the desired part $|I\ (0)|$ decreases and the undesired part $|I(l-k)|$ increases. The undesired ICI degrades the performance of the system. It is not possible to make reliable data decisions unless the ICI powers of OFDM system are minimized. Thus, an accurate and efficient Inter-Carrier Interference (ICI) reduction procedure is essential to demodulate the received data. Several methods have been presented to reduce ICI, including windowing at the receiver [Muschallik, 1996; Müller-Weinfurtner, 2001; Song & Leung, 2005], the use of pulse shaping [Tan & Beaulieu, 2004; Mourad, 2006; Maham & Hjørungnes, 2007], self-cancellation schemes [Zhao & Haggman, 2001], and frequency domain equalization. The next part of this chapter introduces these techniques.

3. ICI reduction techniques

In the OFDM systems, *N* subcarriers are used for data transmission of *N* symbols $\{D_{i,0}, D_{i,1}, \dots D_{i,N-1}\}$. By using the IFFT operation for the data modulation, rectangular pulse shaping filter is implicitly applied. Thus, the spectrum of each individual subcarrier equals a sinc-function defined as $sinc(x) \; = \; sin\;(\pi x)\;/\;\pi x$ and is given by:

$$S_k(z) = D_{i,k} \times sinc(z - z_{k)}\,, \qquad k = 0,1, \dots, N-1 \tag{13}$$

where $z \in \mathbb{R}$ represents the frequency *f* shifted to the carrier frequency of the OFDM system f_c and normalized to the sampling frequency $1/T_s$. The normalized frequency is given by:

$$z = (f - f_c) \times T_s \tag{14}$$

Accordingly, $z_k = (f_k - f_c) \times T_s$ is defined as the normalized center frequency of the *k*-th subcarrier with f_k representing the center frequency of the *k*-th subcarrier. The spectrum of the transmitted OFDM symbol is the superposition of the spectra of all individual subcarriers:

$$S(z) = \sum_{k=0}^{N-1} S_k\,(z) \tag{15}$$

The sidelobe power of this sum signal and also the sidelobe power of each subcarrier spectrum only decays with $1/\mathrm{z}^2$ resulting in a high interference caused by the high sidelobes of the adjoining carriers on a particular subcarrier.

Here, some techniques introduced to reduce the power of the interfering components.

3.1 Pulse shaping

As we have seen in the OFDM spectrum each carrier consists of a main lobe followed by a number of sidelobes with reducing amplitudes. As long as orthogonality is maintained, there is no interference among the carriers because at the peak of the every carrier, there exists a spectral null. That is at that point that the component of all other carriers is zero. Hence the individual carrier is easily separated.

In the presence of the frequency offset the orthogonality is lost because the spectral null does not coincide to the peak of the individual carriers. So some power of the sidelobes exists at the centre of the individual carriers which is called ICI power. The ICI power will go on increasing as the frequency offset increases. The purpose of pulse shaping is to reduce the sidelobes which leads to the significant decrease in the ICI power.

In a simple OFDM system, symbols are performed using an N-FFT function. This implies that the received signal *r(k)* is shaped in the time domain by a rectangular pulse function. One possible countermeasure to overcome the interference is making the PDS of the OFDM modulated subcarriers ($S_n(z)$) go down more rapidly by shaping the transmit signal of the OFDM subcarriers. This makes the amplitude go smoothly to zero at the symbol boundaries. The N-subcarrier OFDM block with pulse-shaping is expressed as:

$$s(t) = e^{j2\pi f_c t} \sum_{k=0}^{N-1} D_k p(t) e^{j2\pi f_k t} \tag{16}$$

where *p(t)* is the pulse shaping function. The transmitted symbol D_k is assumed to have zero mean and normalized average symbol energy. Also we assume that all data symbols are uncorrelated, i.e.:

$$E[D_k D_m^*] = \begin{cases} 1, & k = m, \ k,m = 0,1,\dots,N-1 \\ 0, & k \neq m, \ k,m = 0,1,\dots,N-1 \end{cases} \tag{17}$$

where D_k^* is the complex conjugate of D_k. To ensure the subcarrier orthogonality, which is very important for OFDM systems the equation below has to be satisfied:

$$f_k - f_m = \frac{k-m}{T_s}, \qquad k,m = 0,1,\dots,N-1 \tag{18}$$

In the receiver block, the received signal can be expressed as:

$$r(t) \ = s(t) \otimes h(t) \ + w(t) \tag{19}$$

where ⊗ denotes convolution and h(t) is the channel impulse response. In (19), w(t) is the additive white Gaussian noise process with zero mean and variance $\mathrm{N_0}/2$ per dimension. For this work we assume that the channel is ideal, i.e., h(t) = δ(t) in order to investigate the effect of the frequency offset only on the ICI performance. At the receiver, the received signal $\mathrm{r}^{'}$(t) becomes:

$$r'(t) = e^{j2\pi\Delta f t+\theta} \sum_{k=0}^{N-1} D_k p(t) e^{j2\pi f_k t} + w(t) e^{j(2\pi(-f_c+\Delta f)t+\theta)} \tag{20}$$

Where θ is the phase error and Δf is the carrier frequency offset between transmitter and receiver oscillators. For the transmitted symbol D_m , the decision variable is given as

$$\widehat{D}_m = \int_{-\infty}^{\infty} r^{'}(t) e^{-j2\pi f_m t} dt \tag{21}$$

By using (18) and (21), the decision variable $\widehat{D}_m$ can be expressed as

$$\widehat{D}_m = \left(D_m P(-\Delta f) + \sum_{\substack{k=0 \\ k\neq m}}^{N-1} D_k P(\tfrac{m-k}{T_s} - \Delta f) \right) e^{j\theta} + w_m \quad ,m \ = \ 0,\dots,N-1 \tag{22}$$

where *P(f)* is the Fourier transform of *p(t)* and w_m is the independent white Gaussian noise component. In (22), the first term contains the desired signal component and the second term represents the ICI component. With respect to (18), *P(f)* should have spectral nulls at the frequencies $\pm(1/T_s)$, $\pm(2/T_s)$,... to ensure subcarrier orthogonality. Then, there exists no ICI term if Δf and θ are zero.

The power of the desired signal can be calculated as [Tan & Beaulieu, 2004; Mourad, 2006; Kumbasar & Kucur, 2007]:

$$\sigma_m^2 = E[D_m P(-\Delta f) D_m^* P(-\Delta f)^*] = E[D_m D_m^*]|P(\Delta f)|^2 = |P(\Delta f)|^2 \tag{23}$$

The power of the ICI can be stated as:

$$\sigma_{ICI_m}^2 = \sum_{\substack{k=0 \\ k\neq m}}^{N-1} \sum_{\substack{n=0 \\ n\neq m}}^{N-1} D_n D_k^* P(\frac{k-m}{T_s} + \Delta f) P(\frac{n-m}{T_s} + \Delta f)^* \tag{24}$$

The average ICI power across different sequences can be calculated as:

$$\overline{\sigma_{ICI}^2} = E\left[\sigma_{ICI_m}^2\right] = \sum_{\substack{k=0 \\ k\neq m}}^{N-1} \left|P(\frac{k-m}{T_0} + \Delta f)\right|^2 \tag{25}$$

As seen in (25) the average ICI power depends on the number of the subcarriers and *P(f)* at frequencies: $\left(\frac{k-m}{T_s} + \Delta f\right),\ \ k \neq m,\ k = 0,1,\ldots,N-1$

The system ICI power level can be evaluated by using the CIR (Carrier-to-Interference power Ratio). While deriving the theoretical CIR expression, the additive noise is omitted.

By using (23) and (25), the CIR can be derived as [Tan & Beaulieu, 2004; Mourad, 2006; Kumbasar & Kucur, 2007]:

$$CIR = \frac{|P(\Delta f)|^2}{\sum_{\substack{k=0 \\ k\neq m}}^{N-1} \left|P(\frac{k-m}{T_s} + \Delta f)\right|^2} \tag{26}$$

Therefore, the CIR of the OFDM systems only depends approximately on the normalized frequency offset. A commonly used pulse shaping function is the raised cosine function that is defined by:

$$g(t) = \begin{cases} \frac{1}{2} + \frac{1}{2}cos\left(\pi + \frac{\pi t}{\alpha T_s}\right), & for\ 0 \leq t < \alpha T_s \\ 1, & for\ \alpha T_s \leq t < T_s \\ \frac{1}{2} + \frac{1}{2}cos\left(\frac{\pi(t-T_s)}{\alpha T_s}\right), & for\ T_s \leq t < (1+\alpha\beta)T_s \end{cases} \tag{27}$$

where α denotes the rolloff factor and the symbol interval T_s is shorter than the total symbol duration *(1 + α)* T_s because adjacent symbols are allowed to partially overlap in the rolloff region. Simulation shows that the benefit of the raised cosine function with respect to the ICI reduction is fairly low.

A number of pulse shaping functions such as Rectangular pulse (REC), Raised Cosine pulse (RC), Better Than Raised Cosine pulse (BTRC), Sinc Power pulse (SP) and Improved Sinc Power pulse (ISP) have been introduced for ICI power reduction. Their Fourier transforms are given, respectively as [Kumbasar & Kucur, 2007]:

$$P_{REC}(f) = sinc(fT_s), \tag{28}$$

$$P_{RC}(f) = sinc(fT_s)\frac{cos\ (\pi\alpha fT_s)}{1-(2\alpha fT_s)^2}, \tag{29}$$

$$P_{BTRC}(f) = sinc(fT_s)\frac{[2\beta fT_s\, sin(\pi\alpha fT_s) + 2\ cos(\pi\alpha fT_s) - 1]}{1+(\beta fT_s)^2}, \tag{30}$$

$$P_{SP}(f) = sinc^n(fT_s), \tag{31}$$

$$P_{ISP}(f) = exp\,\{-a(fT_s)^2\}sinc^n(fT_s), \tag{32}$$

where α ($0 \leq \alpha \leq 1$) is the rolloff factor, $\beta = \pi\alpha/ln\ 2$, *a* is a design parameter to adjust the amplitude and *n* is the degree of the sinc function.

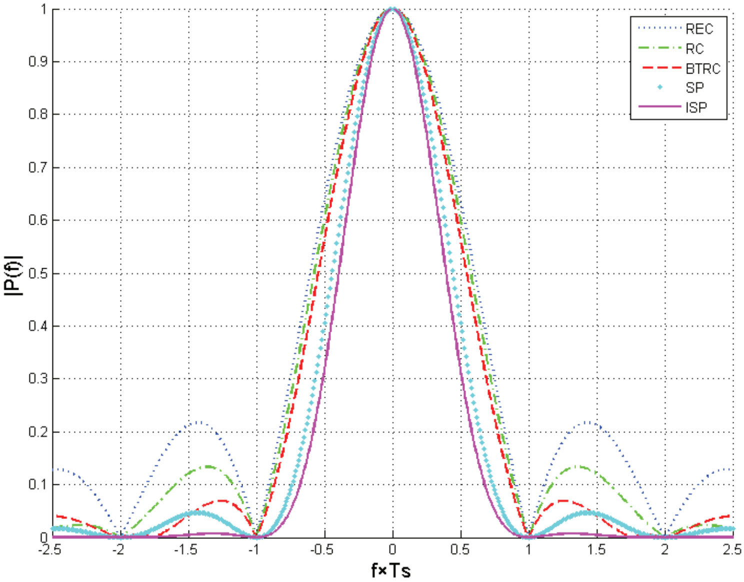


Fig. 5. Comparison of REC, RC, BTRC, SP, and ISP spectrums

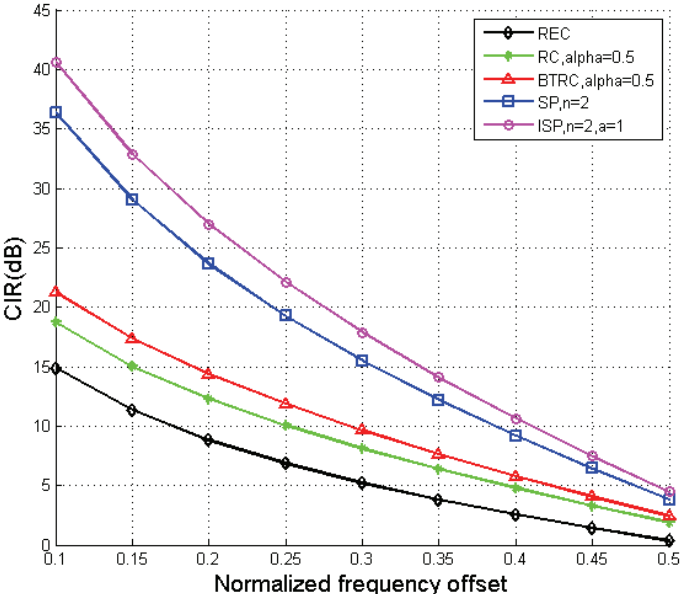


Fig. 6. CIR performance for different pulse shapes

The purpose of pulse shaping is to increase the width of the main lobe and/or reduce the amplitude of sidelobes, as the sidelobe contains the ICI power.

REC, RC, BTRC, SP, and ISP pulse shapes are depicted in Figure 5 for *a=1*, *n=2*, and *α =0.5*. SP pulse shape has the highest amplitude in the main lobe, but at the sidelobes it has lower amplitude than BTRC. This property provides better CIR performance than that of BTRC as shown in [Mourad, 2006]. As seen in this figure the amplitude of ISP pulse shape is the lowest at all frequencies. This property of ISP pulse shape will provide better CIR performance than those of the other pulse shapes as shown in Figure 6 [Kumbasar & Kucur, 2007].

Figure 5 shows that the sidelobe is maximum for rectangular pulse and minimum for ISP pulse shapes. This property of ISP pulse shape will provide better performance in terms of ICI reduction than those of the other pulse shapes. Figure 7 compares the amount of ICI for different pulse shapes.

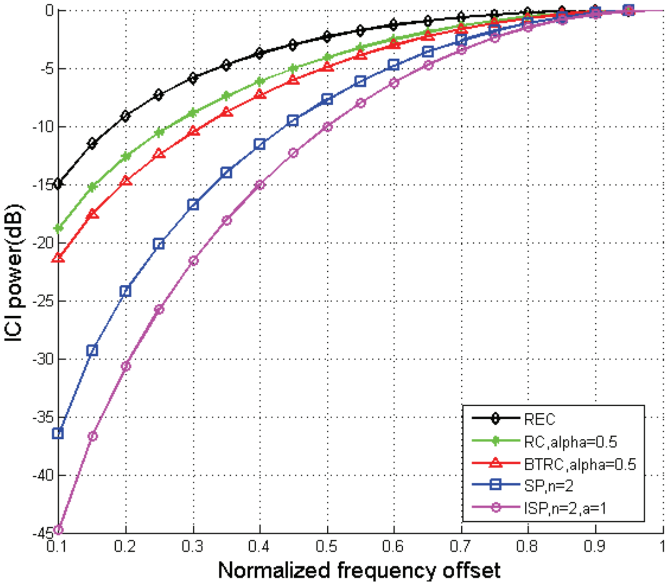


Fig. 7. ICI comparison for different pulse shapes

3.2 ICI self-cancellation methods

In single carrier communication system, phase noise basically produces the rotation of signal constellation. However, in multi-carrier OFDM system, OFDM system is very vulnerable to the phase noise or frequency offset. The serious inter-carrier interference (ICI) component results from the phase noise. The orthogonal characteristics between subcarriers are easily broken down by this ICI so that system performance may be considerably degraded.

There have been many previous works in the field of ICI self-cancellation methods [Ryu et al., 2005; Moghaddam & Falahati, 2007]. Among them convolution coding method, data-conversion method and data-conjugate method stand out.

3.2.1 ICI self-cancelling basis

As it can be seen in eq. 12 the difference between the ICI coefficients of the two consecutive subcarriers are very small. This makes the basis of ICI self cancellation. Here one data symbol is not modulated into one subcarrier, rather at least into two consecutive subcarriers. This is the ICI cancellation idea in this method.

As shown in figure 7 for the majority of *l-k* values, the difference between $I(l-k)$ and $I(l-k+1)$ is very small. Therefore, if a data pair *(a,-a)* is modulated onto two adjacent subcarriers $(l, l+1)$, then the ICI signals generated by the subcarrier will be cancelled out significantly by the ICI generated by subcarrier *l+1* [Zhao & Haggman, 1996, 2001].

Assume that the transmitted symbols are constrained so that $D_1 = -D_0,\ D_3 = -D_2 \ldots D_{N-1} = -D_{N-2}$, then the received signal on subcarrier k considering that the channel coefficients are the same in two adjacent subcarriers becomes:

$$Y_k' = \sum_{\substack{l=0 \\ l=even}}^{N-2} D_l H_l [I(l-k) - I(l-k+1)] + w_k \tag{33}$$

In such a case, the ICI coefficient is denoted as:

$$I'(l-k) = I(l-k) - I(l-k+1) \tag{34}$$

For most of the $l-k$ values, it is found that $|I'\ (l-k)| \ << \ |I\ (l-k)|$.

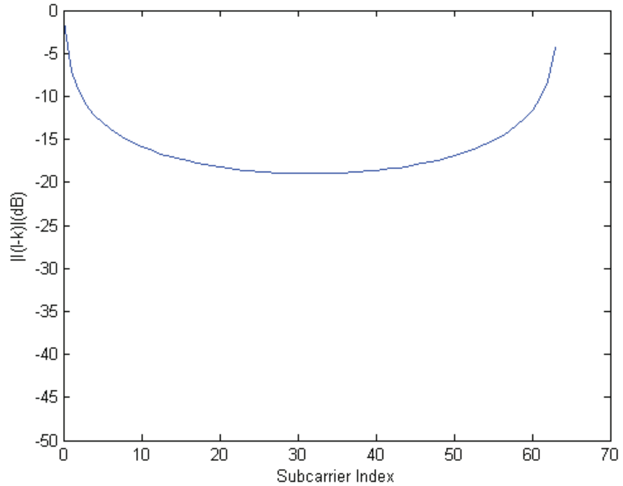


Fig. 7. ICI coefficient versus subcarrier index; N=64

For further reduction of ICI, ICI cancelling demodulation is done. The demodulation is suggested to work in such a way that each signal at the *k+1*-th subcarrier (now *k* denotes even number) is multiplied by *-1* and then summed with the one at the *k*-th subcarrier. Then the resultant data sequence is used for making symbol decision. It can be represented as:

$$Y_k^{"} = Y_k' - Y_{k+1}' = \sum_{\substack{l=0 \\ l=even}}^{N-2} D_l H_l [-I(l-k-1) + 2I(l-k) - I(l-k+1)] + w_k - w_{k+1} \tag{35}$$

The corresponding ICI coefficient then becomes:

$$I"(l-k) = -I(l-k-1) + 2I(l-k) - I(l-k+1) \tag{36}$$

Figure 8 shows the amplitude comparison of $|I(l-k)|$, $|I'(l-k)|$ and $|I"(l-k)|$ for *N=64* and ε = 0.3. For the majority of *l-k* values, $|I'(l-k)|$ is much smaller than $|I\ (l-k)|$, and the $|I"(l-k)|$ is even smaller than $|I'(l-k)|$. Thus, the ICI signals become smaller when applying ICI cancelling modulation. On the other hand, the ICI cancelling demodulation can further reduce the residual ICI in the received signals. This combined ICI cancelling modulation and demodulation method is called the ICI self-cancellation scheme.

Due to the repetition coding, the bandwidth efficiency of the ICI self-cancellation scheme is reduced by half. To fulfill the demanded bandwidth efficiency, it is natural to use a larger signal alphabet size. For example, using 4PSK modulation together with the ICI self-cancellation scheme can provide the same bandwidth efficiency as standard OFDM systems (1 bit/Hz/s).

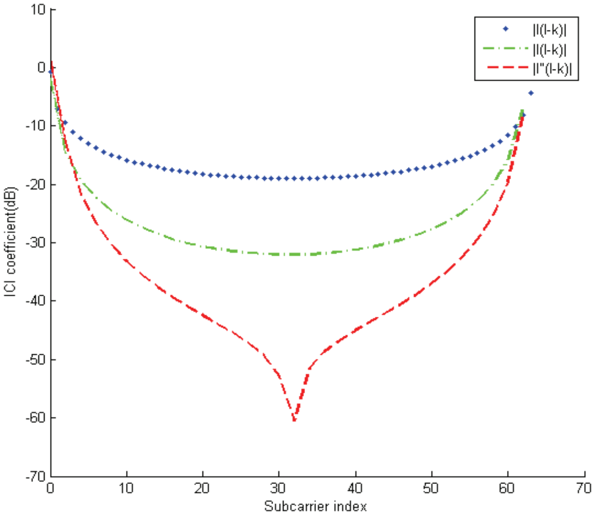


Fig. 8. Amplitude comparison of $|I\ (l-k)|$, $|I'\ (l-k)|$ and $|I"\ (l-k)|$

3.2.1.1 Data-conjugate method

In an OFDM system using data-conjugate method, the information data pass through the serial to parallel converter and become parallel data streams of *N/2* branch. Then, they are converted into *N* branch parallel data by the data-conjugate method. The conversion process is as follows. After serial to parallel converter, the parallel data streams are remapped as the form of $D'_{2k} = D_k$, $D'_{2k+1} = -D^*_k$, (k = 0, … , $N/2$-1). Here, D_k is the information data to the *k*-th branch before data-conjugate conversion, and D'_{2k} is the information data to the 2*k*-th branch after ICI cancellation mapping. Likewise, every information data is mapped into a pair of adjacent sub-carriers by data-conjugate method, so the *N/2* branch data are extended to map onto the *N* sub-carries.

The original data can be recovered from the simple relation of $Z'_k = (Y_{2k} - Y^*_{2k+1})/2$. Here, Y_{2k} is the *2k*-th sub-carrier data, Z'_k is the *k*-th branch information data after de-mapping. Finally, the information data can be found through the detection process. The complex base-band OFDM signal after data conjugate mapping is as follows.

$$d(n) = \sum_{i=0}^{N-1} D_i'.e^{j\frac{2\pi}{N}in} = \sum_{k=0}^{\frac{N}{2}-1} \left[D_k.e^{j\frac{2\pi}{N}2kn} - D_k^*.e^{j\frac{2\pi}{N}(2k+1)n}\right], \qquad for\ 0 \le n < N \tag{37}$$

where, *N* is the total number of sub-carriers, *D_k* is data symbol for the *k*-th parallel branch and D_k' is the *i*–th sub-carrier data symbol after data-conjugate mapping. *d(n)* is corrupted by the phase noise in the transmitter (TX) local oscillator. Furthermore, the received signal is influenced by the phase noise of receiver (RX) local oscillator. So, it is expressed as:

$$r(t) = \left\{\left(s(t).e^{j\theta_{TX}(t)}\right) \otimes h(t) + w(t)\right\}.e^{j\theta_{RX}(t)} \tag{38}$$

where *s(t)* is the transmitted signal, *w(t)* is the white Gaussian noise and *h(t)* is the channel impulse response. $\theta_{TX}(t)$ and $\theta_{RX}(t)$are the time varying phase noise processes generated in the transceiver oscillators. Here, it is assumed that, $\theta_{TX}(t) = \theta_{RX}(t) = \theta(t)$ and $\theta_{tot}(t) = \theta_{TX}(t) + \theta_{RX}(t)$ for simple analysis. In the original OFDM system without ICI self-cancellation method, the *k*-th sub-carrier signal after FFT can be written as:

$$Y_k = \frac{1}{N}\sum_{l=0}^{N-1} D_l.H_l \sum_{m=0}^{N-1} e^{j\left(\frac{2\pi}{N}(l-k)m+\theta_{tot}(m)\right)} + w_k \tag{39}$$

In the data-conjugate method, the sub-carrier data is mapped in the form of $D_{2k}^{'} = D_k, D_{2k+1}^{'} = -D_k^*$. Therefore, the 2*k*-th sub-carrier data after FFT in the receiver is arranged as:

$$Y_{2k} = \sum_{l=0}^{\frac{N}{2}-1} [D_l H_{2l} Q_{2l-2k} - D_l^* H_{2l+1} Q_{2l+1-2k}] + w_{2k} \tag{40}$$

$$Q_L = \frac{1}{N}\sum_{m=0}^{N-1} e^{j\left(\frac{2\pi}{N}Lm+\theta_{tot}(m)\right)} \tag{41}$$

w_{2k} is a sampled FFT version of the complex AWGN multiplied by the phase noise of RX local oscillator, and random phase noise process $\theta_{tot}(m)$ is equal to $\theta_{TX}(m) + \theta_{RX}(m)$.
Similarly, the *2k+1*-th sub-carrier signal is expressed as:

$$Y_{2k+1} = \sum_{l=0}^{\frac{N}{2}-1} [D_l H_{2l} Q_{2l-2k-1} - D_l^* H_{2l+1} Q_{2l-2k}] + w_{2k+1} \tag{42}$$

In the (40) and (42), $l = k$ corresponds to the original signal with CPE, and $l \neq k$ corresponds to the ICI component. In the receiver, the decision variable Z_k' of the *k*-th symbol is found from the difference of adjacent sub-carrier signals affected by phase noise. That is,

$$\begin{aligned} Z_k^{'} &= \frac{(Y_{2k} - Y_{2k+1}^*)}{2} = \frac{1}{2}X_k(H_{2k}Q_0 + H_{2k+1}^*Q_0^*) - \frac{1}{2}D_k^*(H_{2k+1}Q_1 + H_{2l}^*Q_{-1}^*) \\ &+ \frac{1}{2}\sum_{\substack{l=0 \\ l\neq k}}^{\frac{N}{2}-1}\{D_l[H_{2l}Q_{2l-2k} + H_{2l+1}^*Q_{2l-2k}^*] - D_l^*[H_{2l+1}Q_{2l+1-2k} + H_{2l}^*Q_{2l-2k-1}^*]\} + w_k \end{aligned} \tag{43}$$

where $w_k = (1/2)(w_{2k} - w_{2k+1}^*)$ is the AWGN of the *k*–th parallel branch data in the receiver. When channel is flat, frequency response of channel $\{H_k\}$ equals 1. Z'_k is as follows.

$$Z_k' = D_k + \frac{1}{2}\sum_{\substack{l=0\\ l\neq k}}^{\frac{N}{2}-1}\{D_l[Q_{2l-2k} + Q_{2l-2k}^*] - D_l^*[Q_{2l+1-2k} + Q_{2l-2k-1}^*]\} + w_k \tag{44}$$

3.3 CPE, ICI and CIR analysis

A. Original OFDM

In the original OFDM, the k-th sub-carrier signal after FFT is as follows:

$$Y_k = D_k Q_0 + \sum_{\substack{l=0\\ l\neq k}}^{N-1} D_l . Q_{l-k} + w_k \tag{45}$$

The received desired signal power on the k-th sub-carrier is:

$$E[|Y_{k1}|^2] = E[|D_k Q_0|^2] \tag{46}$$

ICI power is:

$$E[|Y_{k2}|^2] = E\left[\left|\sum_{\substack{l=0\\ l\neq k}}^{N-1} D_l Q_{l-k}\right|^2\right] \tag{47}$$

Transmitted signal is supposed to have zero mean and statistically independence. So, the CIR of the original OFDM transmission method is as follows:

$$CIR = \frac{|Q_0|^2}{\sum_{\substack{l=0\\ l\neq k}}^{N-1}|Q_{l-k}|^2} = \frac{|Q_0|^2}{\sum_{l=1}^{N-1}|Q_l|^2} \tag{48}$$

B. Data-conversion method

In the data-conversion ICI self-cancellation method, the data are remapped in the form of $D_{2k}^{'} = D_k$, $D_{2k+1}^{'} = -D_k$.
So, the desired signal is recovered in the receiver as follows:

$$\begin{aligned} Z_k' = \frac{(Y_{2k} - Y_{2k+1})}{2} &= D_k + \frac{1}{2}D_k[-Q_{-1} + 2(Q_0 - 1) - Q_1] \\ &+ \frac{1}{2}\sum_{\substack{l=0\\ l\neq k}}^{\frac{N}{2}-1} D_l[-Q_{2l-2k-1} + Q_{2l-2k} - Q_{2l-2k+1}] + w_k \end{aligned} \tag{49}$$

CPE is as follows:

$$CPE = \frac{j2D_k}{N}\sum_{m=0}^{N-1} sin^2\left(\frac{\pi m}{N}\right)\theta_{tot}(m) \tag{50}$$

ICI component of the k-th sub-carrier is as follows:

$$ICI = \frac{2j}{N}\sum_{\substack{l=0\\ l\neq k}}^{\frac{N}{2}-1} D_l \sum_{m=0}^{N-1} sin^2\left(\frac{\pi m}{N}\right).exp\left(\frac{j4\pi m(l-k)}{N}\right)\theta_{tot}(m) \tag{51}$$

So

$$CIR = \frac{|-Q_{-1}+2Q_0-Q_1|^2}{\sum_{\substack{l=0\\ l\neq k}}^{\frac{N}{2}-1}|-Q_{2l-2k-1}+2Q_{2l-2k}-Q_{2l-2k+1}|^2} = \frac{|-Q_{-1}+2Q_0-Q_1|^2}{\sum_{l=1}^{\frac{N}{2}-1}|-Q_{2l-1}+2Q_{2l}-Q_{2l+1}|^2} \tag{52}$$

C. Data-conjugate method

In the data conjugate method, the decision variable can be written as follows:

$$Z'_k = D_k + \frac{1}{2}\sum_{\substack{l=0\\ l\neq k}}^{\frac{N}{2}-1}\{D_l[Q_{2l-2k}+Q^*_{2l-2k}] - D^*_l[Q_{2l+1-2k}+Q^*_{2l-2k-1}]\} + w_k \tag{53}$$

Through the same calculation, CPE, ICI and CIR of the data conjugate method are found.

$$CPE = 0 \tag{54}$$

The fact CPE is zero is completely different from the data conversion method whose CPE is not zero like (14).
Then, ICI of data conjugate method is:

$$ICI = \frac{1}{N}\sum_{\substack{l=0\\ l\neq k}}^{\frac{N}{2}-1}\sum_{m=0}^{N-1} sin\left(\frac{4\pi m(l-k)}{N}\right).[D^*_l.e^{j\frac{2\pi}{N}m} - D_l].\theta_{tot}(m) \tag{55}$$

The above term is the summation of the signal of the other sub-carriers multiplied by some complex number resulted from an average of phase noise with spectral shift. This component is added into the *k*-th branch data Z'_k. It may break down the orthogonalities between sub-carriers. So, CIR is:

$$CIR = \frac{4}{\sum_{l=2}^{\frac{N}{2}-1}\left[\left|Q_{2l}+Q^*_{2l}\right|^2 + \left|Q_{2l+1}+Q^*_{2l-1}\right|^2\right]} \tag{56}$$

4. Conclusion

OFDM has been widely used in communication systems to meet the demand for increasing data rates. It is robust over multipath fading channels and results in significant reduction of the transceiver complexity. However, one of its disadvantages is sensitivity to carrier frequency offset which causes attenuation, rotation of subcarriers, and inter-carrier interference (ICI). The ICI is due to frequency offset or may be caused by phase noise.
The undesired ICI degrades the signal heavily and hence degrades the performance of the system. So, ICI mitigation techniques are essential to improve the performance of an OFDM system in an environment which induces frequency offset error in the transmitted signal. In this chapter, the performance of OFDM system in the presence of frequency offset is

analyzed. This chapter investigates different ICI reduction schemes for combating the impact of ICI on OFDM systems. A number of pulse shaping functions are considered for ICI power reduction and the performance of these functions is evaluated and compared using the parameters such as ICI power and CIR. Simulation results show that ISP pulse shapes provides better performance in terms of CIR and ICI reduction as compared to the conventional pulse shapes.

Another ICI reduction method which is described in this chapter is the ICI self cancellation method which does not require very complex hardware or software for implementation. However, it is not bandwidth efficient as there is a redundancy of 2 for each carrier. Among different ICI self cancellation methods, the data-conjugate method shows the best performances compared with the original OFDM, and the data-conversion method since it makes CPE to be zero along with its role in significant reduction of ICI.

5. References

Robertson, P. & Kaiser, S. (1995). Analysis of the effects of phase-noise in orthogonal frequency division multiplex (OFDM) systems, *Proceedings of the IEEE International Conference on Communications,* vol. 3, (Seattle, USA), pp. 1652–1657, June 1995.

Zhao, Y. & Haggman, S.G. (2001). Intercarrier interference self-cancellation scheme for OFDM mobile communication systems, *IEEE Transaction on Communication.* pp. 1185–1191.

Muschallik, C. (1996). Improving an OFDM reception using an adaptive Nyquist windowing, *IEEE Transaction Consum*. Electron. 42 (3) (1996) 259–269.

Müller-Weinfurtner, S.H. (2001). Optimum Nyquist windowing in OFDM receivers, *IEEE Trans. Commun.* 49 (3) (2001) 417–420.

Song, R. & Leung, S.-H. (2005). A novel OFDM receiver with second order polynomial Nyquist window function, *IEEE Communication Letter.* 9 (5) (2005) 391–393.

Tan, P. & Beaulieu, N.C. (2004). Reduced ICI in OFDM systems using the better than raised-cosine pulse, *IEEE Communication Letter* 8 (3) (2004) 135–137.

Mourad, H.M. (2006). Reducing ICI in OFDM systems using a proposed pulse shape, *Wireless Person. Commun.* 40 (2006) 41–48.

Kumbasar, V. & Kucur, O. (2007). ICI reduction in OFDM systems by using improved sinc power pulse, *ELSEVIER Digital Signal Processing* 17 (2007) 997-1006

Zhao, Y. & Häggman, S.-G. (1996). Sensitivity to Doppler shift and carrier frequency errors in OFDM systems—The consequences and solutions, *Proceeding of IEEE 46th Vehicular Technology Conference,* Atlanta, GA, Apr. 28–May 1, 1996, pp. 1564–1568.

Ryu, H. G.; Li, Y. & Park, J. S. (2005). An Improved ICI Reduction Method in OFDM Communication System, *IEEE Transaction on Broadcasting*, Vol. 51, No. 3, September 2005.

Mohapatra, S. & Das, S. (2009). Performance Enhancement of OFDM System with ICI Reduction Technique, *Proceeding of the World Congress on Engineering 2009*, Vol. 1, WCE 2009, London, U.K.

Moghaddam, N. & Falahati, A. (2007). An Improved ICI Reduction Method in OFDM Communication System in Presence of Phase Noise, *the 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07)*

Kumar, R. & Malarvizhi, S. (2006). Reduction of Intercarrier Interference in OFDM Systems.

Maham, B. & Hjørungnes, A. (2007). ICI Reduction in OFDM by Using Maximally Flat Windowing, *IEEE International Conference on Signal Processing and Communications (ICSPC 2007),* Dubai, United Arab Emirates (UAE).

4

Multiple Antenna Techniques

Han-Kui Chang, Meng-Lin Ku, Li-Wen Huang and Jia-Chin Lin
Department of Communication Engineering, National Central University, Taiwan, R.O.C.

1. Introduction

Recent developed information theory results have demonstrated the enormous potential to increase system capacity by exploiting multiple antennas. Combining multiple antennas with orthogonal frequency division multiplexing (OFDM) is regarded as a very attractive solution for the next-generation wireless communications to effectively enhance service quality over multipath fading channels at affordable transceiver complexity. In this regard, multiple antennas, or called multiple-input multiple-output (MIMO) systems, have emerged as an essential technique for the next-generation wireless communications. In general, an MIMO system has capability to offer three types of antenna gains: diversity gains, multiplexing gains and beamforming gains. A wide variety of multiple antennas schemes have been investigated to achieve these gains, while some combo schemes can make trade-offs among these three types of gains. In this chapter, an overview of multiple antenna techniques developed in the past decade, as well as their transceiver architecture designs, is introduced. The first part of this chapter covers three kinds of diversity schemes: maximum ratio combining (MRC), space-time coding (STC), and maximum ratio transmission (MRT), which are commonly used to combat channel fading and to improve signal quality with or without channel knowledge at the transmitter or receiver. The second part concentrates on spatial multiplexing to increase data rate by simultaneously transmitting multiple data streams without additional bandwidth or power expenditure. Several basic receiver architectures for handling inter-antenna interference, including zero-forcing (ZF), minimum mean square error (MMSE), interference cancellation, etc., are then introduced. The third part of this chapter introduces antenna beamforming techniques to increase signal-to-interference plus noise ratio (SINR) by coherently combining signals with different phase and amplitude at the transmitter or receiver, also known as transmit beamforming or receive beamforming. Another benefit of adopting beamforming is to facilitate multiuser accesses in spatial domain and effectively control multiuser interference. The optimal designs of these beamforming schemes are also presented in this chapter.

2. Diversity techniques

Diversity techniques have been widely adopted in modern communications to overcome multipath fading, which allows for enhancing the reliability of signal reception without sacrificing additional transmission power and bandwidth (Rappaport, 2002; Simon & Alouini, 1999). The basic idea of diversity is that multiple replicas of transmitted signals which carry the same information, but experience independent or small correlated fading, are available at the receiver. In fading channels, some samples are severely faded, while others are less attenuated; hence, in statistics, the probability of the signal strength of all samples being simultaneously below a given level becomes small, as compared with the case without applying diversity techniques. Consequently, we can overwhelm the channel fading by imposing an appropriate selection or combination of various samples, so as to dramatically improve the signal quality. Based on signal processing domains to obtain diversity gains, diversity techniques can be classified into time, frequency and space diversity. Here, we focus on space diversity techniques where multiple antennas are deployed at the transmitter or receiver sides. One category of space diversity schemes is to combine multiple signal replicas at the receiver, which is termed as receive diversity. The other category is to use multiple antennas at the transmitter, and this kind of diversity schemes is called transmit diversity (Giannakis et al., 2006).

In this section, we first present various receive diversity schemes, including selection combining, switch combining, equal-gain combining (EGC), and MRC. The well-known Alamouti's transmit diversity scheme using two transmit antennas and one receive antenna is then introduced. The generalized case using two transmit antennas and multiple receive antennas is shown as well. Subsequently, space-time block codes (STBCs) with the number of transmit antennas larger than two (Tarokh et al., 1999) are presented. Finally, a maximum ratio transmission (MRT) scheme is discussed to simultaneously achieve both transmit and receive diversity gains and maximize the output signal-to-noise ratio (SNR) (Lo, 1999).

2.1 Receive diversity techniques

In cellular systems, receive diversity techniques have been widely applied at base stations for uplink transmission to improve the signal reception quality. This is mainly because base stations can endure larger implementation size, power consumption, and cost. In general, the performance of the receive diversity not only depends on the number of antennas but also the combining methods utilized at the receiver side. According to the implementation complexity and the extent of channel state information required at the receiver, we will introduce four types of combining schemes, including selection combining, switch combining, EGC, and MRC, in the following.

2.1.1 Selection combining

Selection combining is a simple receive diversity combining scheme. Consider a receiver equipped with n_R receive antennas. Fig. 1 depicts the block diagram of the selection combining scheme. The antenna branch with the largest instantaneous SNR is selected to receive signals at every symbol period. In practical, since it is difficult to measure the SNR, one can implement the selection combining scheme by accumulating and averaging the received signal power, consisting of both signal and noise power, for all antenna branches, and selecting one branch with the highest output signal power.

2.1.2 Switch combining

Fig. 2 shows the switch combining diversity scheme. As its name suggested, the receiver scans all the antenna branches and selects a certain branches with the SNR values higher than a preset threshold to receive signals. When the SNR of the selected antenna is dropped down the given threshold due to channel fading, the receiver starts scanning all branches again and switches to other antenna branches. As compared with the selection diversity scheme, the switch diversity scheme exhibits lower performance gain since it does not pick up the branch with the highest instantaneous SNR or received signal power. In spite of this performance loss, it is still very attractive for practical implementation as it does not require to periodically and simultaneously monitor all the antenna branches. Another advantage is that since both the selection and switch diversity schemes do not require any knowledge of channel state information, they are not limited to coherent modulation schemes, but can also be applied for noncoherent modulation schemes.

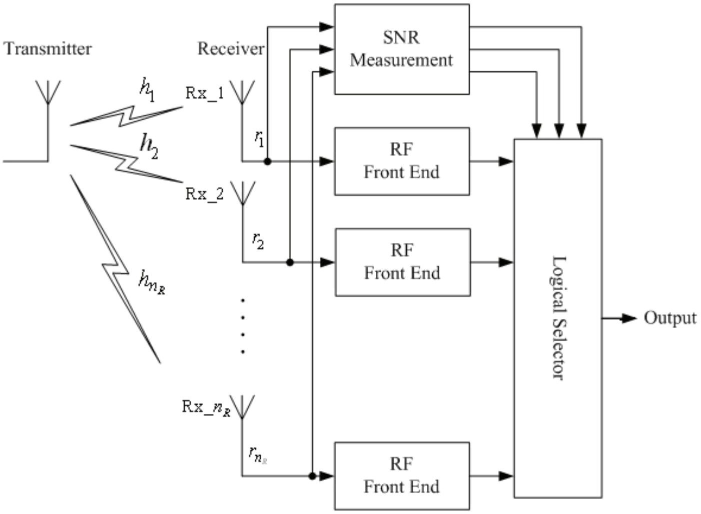


Fig. 1. Block diagram of selection combining scheme

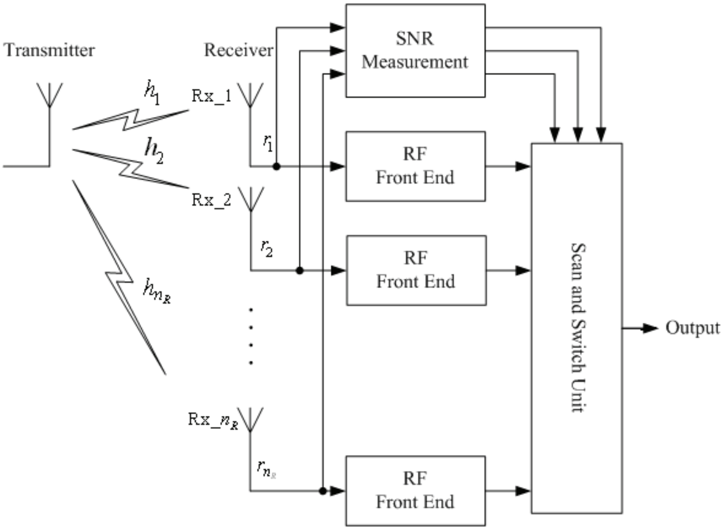


Fig. 2. Block diagram of switch combining scheme

2.1.3 Maximum ratio combining

Fig. 3 shows the block diagram of the MRC scheme. MRC is a linear combining scheme, in which multiple received replicas at the all antenna branches are individually weighted and summed up as an output signal. Since the multiple replicas experience different channel fading gains, the combining scheme can provide diversity gains. In general, there are several ways to determine the weighting factors. Consider a receiver having n_R receive antennas, and the received signals can be expressed as a matrix-vector form as follows:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ \vdots \\ r_{n_R} \end{bmatrix} = \begin{bmatrix} h_1 \\ \vdots \\ h_{n_R} \end{bmatrix} s + \begin{bmatrix} n_1 \\ \vdots \\ n_{n_R} \end{bmatrix} = s\mathbf{h} + \mathbf{n} \tag{11}$$

where r_i , h_i , and n_i are the received signal, channel fading gain, and spatially white noise at the *ith* receive antenna branch, respectively. After linearly combining the received signals, the output signal is given by

$$y = \mathbf{w}^\dagger \mathbf{r} = \mathbf{w}^\dagger (s\mathbf{h} + \mathbf{n}) = s\mathbf{w}^\dagger \mathbf{h} + \mathbf{w}^\dagger \mathbf{n} \tag{2}$$

where $\mathbf{w}$ represents the weighting factors for all antenna branches, and $(\bullet)^\dagger$ is the Hermitian operation. Subsequently, from (2), for a given $\mathbf{h}$, the output SNR is calculated by

$$SNR_o = \frac{E_s \left|\mathbf{w}^\dagger \mathbf{h}\right|^2}{\sigma_n^2 \left\|\mathbf{w}\right\|^2} \tag{3}$$

where E_s and σ_n^2 are the signal power and the noise power, respectively. According to the Cauchy-Schwarz inequality, we have

$$\left|\mathbf{w}^\dagger \mathbf{h}\right|^2 \le \left\|\mathbf{w}\right\|^2 \left\|\mathbf{h}\right\|^2 \tag{4}$$

Hence, the upper bound for the output SNR is given by

$$SNR_o = \frac{E_s \left|\mathbf{w}^\dagger \mathbf{h}\right|^2}{\sigma_n^2 \left\|\mathbf{w}\right\|^2} \le \frac{E_s \left\|\mathbf{h}\right\|^2}{\sigma_n^2} = \left\|\mathbf{h}\right\|^2 SNR_i \tag{5}$$

where $SNR_i = E_s / \sigma_n^2$ is defined as the input SNR. We can further observe that the equality in (5) holds if and only if $\mathbf{w} = \mathbf{h}$, and therefore, the maximum output SNR can be written as

$$SNR_o = \left\|\mathbf{h}\right\|^2 SNR_i \tag{6}$$

The method adopting weighting factors $\mathbf{w} = \mathbf{h}$ is called MRC, as it is capable of maximizing the output SNR with a combining gain of $\left\|\mathbf{h}\right\|^2$. However, the main drawback of the MRC scheme is that it requires the complete knowledge of channel state information, including both amplitude and phase of h_i, to coherently combine all the received signals. Hence, it is not suitable for noncoherent modulation schemes.

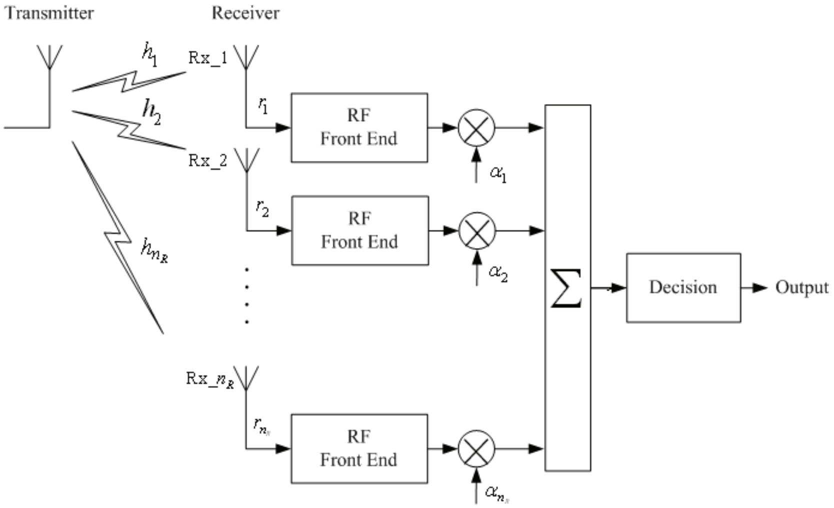


Fig. 3. Block diagram of MRC scheme

2.1.4 Equal gain combining

Equal gain combing is a suboptimal combining scheme, as compared with the MRC scheme. Instead of requiring both the amplitude and phase knowledge of channel state information, it simply needs phase information for each individual channels, and set the amplitude of the weighting factor on each individual antenna branch to be unity. Thus, all multiple received signals are combined in a co-phase manner with an equal gain. The performance of the equal gain combining scheme is only slightly worse than that of the MRC scheme, while its implementation cost is significantly less than that of the MRC scheme.

2.2 Transmit diversity techniques

Although the receive diversity can provide great benefits for uplink transmission, it is difficult to utilize the receive diversity techniques at mobile terminals for downlink transmission. First, it is hard to place more than two antenna elements in a small-size portable mobile device. Second, multiple chains of radio frequency components will increase power consumption and implementation cost. Since mobile devices are usually battery-limited and cost-oriented, it is impractical and uneconomical for using multiple antennas at the mobile terminals to gain diversity gains at forward links. For these reasons, transmit diversity techniques are deemed as a very attractive alternative. Wittneben (Wittneben, 1993) proposed a delay diversity scheme, where replicas of the same symbol are transmitted through multiple antennas at different time slots to impose an artificial multipath. A maximum likelihood sequence estimator (MLSE) or a MMSE equalizer is subsequently used to obtain spatial diversity gains. Another interesting approach is STC, which can be divided into two categories: space-time trellis codes (STTCs) (Tarokh et al., 1998) and STBCs. In the STTC scheme, encoded symbols are simultaneously transmitted through different antennas and decoded using a maximum likelihood (ML) decoder. This scheme combines the benefits of coding gain and diversity gain, while its complexity grows exponentially with the bandwidth efficiency and achievable diversity order. Therefore, it may be not practical or cost-effective for some applications. Alamouti's STC was historically the first STBC to provide two- branch transmit diversity gains for a communication system equipped with two transmit antennas. It has been recognized as a remarkable, but simple, diversity technique, and adopted in a number of next-generation wireless standards, e.g., 3GPP long-term evolution and IEEE 802.16e standards.

In this section, we overview Alamouti's transmit diversity technique. We focus on both encoding and decoding algorithms, along with its performance results. Then, we introduce the generalized STBCs with an arbitrary number of transmit antennas to achieve full diversity gains, which are proposed by Vahid Tarokh (Tarokh et al., 1999) based on orthogonal design theory.

2.2.1 Alamouti's space-time encoding

The encoding procedure of Alamouti's Space-time codes for a two-transmit antenna system is depicted in Fig. 4. Assume that data symbols, each of which is mapped from a group of m information bits through an M-ary modulation scheme, are going to be transmitted, where $m = \log_2 M$. Let $\mathbf{C}$ denote the set of constellation points. For each encoding round, the encoder successively takes a pair of two modulated data symbols $x_1 \in \mathbf{C}$ and $x_2 \in \mathbf{C}$ to generate two transmit signal sequences of length two, according to the following space-time encoding matrix:

$$\mathbf{X} = \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix} \tag{7}$$

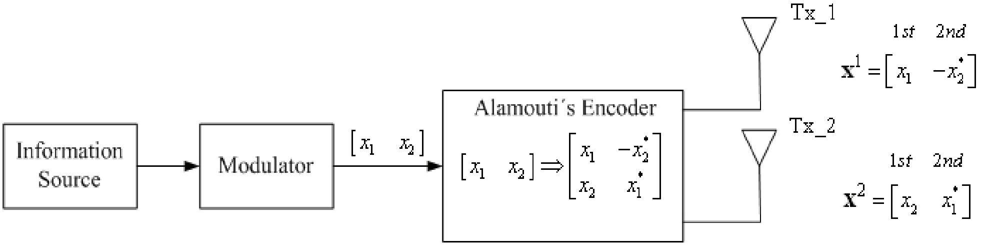

Information Source
Modulator
$[x_1 \quad x_2]$
Alamouti's Encoder
$[x_1 \quad x_2] \Rightarrow \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix}$
Tx_1
1st 2nd
$\mathbf{x}^1 = [x_1 \quad -x_2^*]$
Tx_2
1st 2nd
$\mathbf{x}^2 = [x_2 \quad x_1^*]$

Fig. 4. Block diagram of Alamouti's space-time encoder

The Alamouti's STC is a two-dimensional code, in which the encoder outputs are transmitted within two consecutive time slots over two transmit antennas. During the first time slot, two signals x_1 and x_2 are transmitted simultaneously from antenna one and antenna two, respectively. Similarly, in the second time slot, the signal $-x_2^*$ is transmitted from antenna one and the signal x_1^* is from antenna two, where $(\bullet)^*$ denotes the complex conjugate operation. It is clear that the encoding process is accomplished in both spatial and temporal domains. Let us first denote the transmit sequence from antenna one and two by $\mathbf{x}^1$ and $\mathbf{x}^2$, respectively, as

$$\begin{aligned} \mathbf{x}^1 &= \begin{bmatrix} x_1 & -x_2^* \end{bmatrix} \\ \mathbf{x}^2 &= \begin{bmatrix} x_2 & x_1^* \end{bmatrix} \end{aligned} \tag{8}$$

We can observe that these two signal sequences possess the orthogonal property with each other. That is, we have

$$\mathbf{x}^1\left(\mathbf{x}^2\right)^\dagger = x_1x_2^* - x_2^*x_1 = 0 \tag{9}$$

Where $(\bullet)^\dagger$ denotes the Hermitian operation.
In other words, the code matrix, $\mathbf{X}$, satisfies the orthogonal matrix property as follows:

$$\begin{aligned}\mathbf{XX}^\dagger &= \begin{bmatrix} |x_1|^2+|x_2|^2 & 0 \\ 0 & |x_1|^2+|x_2|^2 \end{bmatrix} \\ &= \left(|x_1|^2+|x_2|^2\right)\mathbf{I}_2\end{aligned} \tag{10}$$

where $\mathbf{I}_2$ is a 2×2 identity matrix.

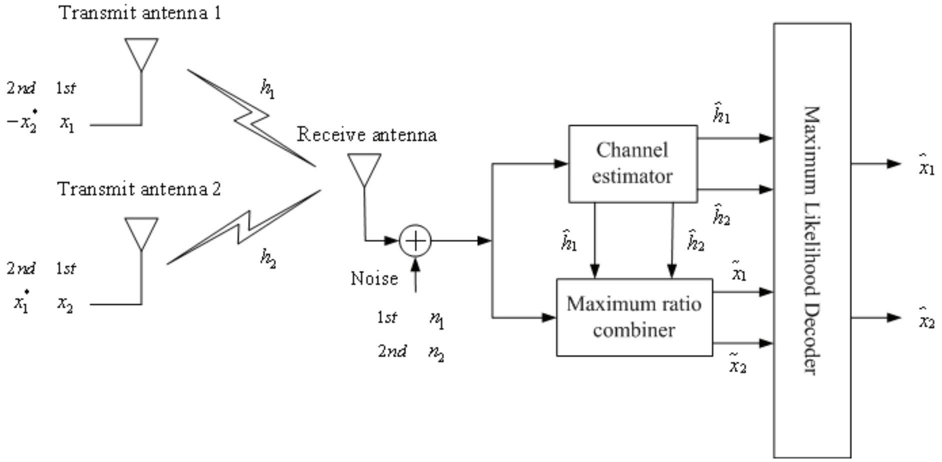


Fig. 5. Block diagram of Alamouti's space-time encoder

Let us assume that there is only one receive antenna deployed at the receiver side. The receiver block diagram for the Alamouti's scheme is shown in Fig. 5. Assume that flat fading channel gains from transmit antenna one and two to the receive antenna at the time slot t are denoted by $h_1(t)$ and $h_2(t)$, respectively. Under the assumption of quasi-static channels, the channel gains across two consecutive symbol periods remain unchanged, and they can be expressed as follows:

$$h_1(t) = h_1(t+T) = h_1 \tag{11}$$

and

$$h_2(t) = h_2(t+T) = h_2 \tag{12}$$

where h_i, for $i=1$ and 2, is a complex constant value corresponding to the channel gain from the transmit antenna i to the receive antenna, and T denotes the symbol period. At the receive antenna, the received signals across two consecutive symbol periods, which are denoted by r_1 and r_2 for time t and $t+T$, are respectively given by

$$r_1 = h_1 x_1 + h_2 x_2 + n_1 \tag{13}$$

and

$$r_2 = -h_1 x_2^* + h_2 x_1^* + n_2 \tag{14}$$

where n_1 and n_2 are independent additive white Gaussian noise with zero mean and variance σ^2. It is noticed here that although we present Alamouti's space-time codes under flat fading channels without concerning the multipath effect, it is straightforward to extend the Alamouti's scheme to the case of multipath channels by using an OFDM technique to transform a frequency selective fading channel into a number of parallel flat fading channels (Ku & Huang, 2006).

2.2.2 Maximum likelihood decoding for Alamouti's scheme

The successful decoding for Alamouti's space-time codes requires the knowledge on channel state information h_1 and h_2 at the receiver side. In general, channel estimation can be performed through the use of some pilot signals which are frequently transmitted from the transmit side (Ku & Huang, 2008; Lin, 2009a, 2009b). Here, we focus on the decoding scheme and assume that channel state information is perfectly estimated and known to the receiver. From the viewpoint of minimum error probability, the decoder intends to choose an optimal pair of constellation points, $\left(\hat{x}_1, \hat{x}_2\right)$, to maximize the a posteriori probability given by the received signals r_1 and r_2. Mathematically, we can express the decoding problem as

$$\left(\hat{x}_1, \hat{x}_2\right) = \arg \max_{(x_1, x_2) \in \mathbf{C}^2} \Pr\left(x_1, x_2 \middle| r_1, r_2\right) \tag{15}$$

where $\mathbf{C}^2$ is the set of all possible candidate symbol pairs $\left(x_1, x_2\right)$, and $\Pr(\bullet)$ is a probability notation. According to the Bayes' theorem, we can further expand (15) as

$$\left(\hat{x}_1, \hat{x}_2\right) = \arg \max_{(x_1, x_2) \in \mathbf{C}^2} \frac{\Pr\left(r_1, r_2 \middle| x_1, x_2\right) \mathrm{P}\left(x_1, x_2\right)}{\mathrm{P}\left(r_1, r_2\right)} \tag{16}$$

By assuming that all the constellation points in $\mathbf{C}^2$ occur with equal prior probabilities and the two symbols of each pair are generated independently, all symbol pairs $\left(x_1, x_2\right)$ are equiprobable. As the decision of the symbol pairs $\left(x_1, x_2\right)$ is irrelevant to the probability of received signals r_1 and r_2, we can rewrite (16) as

$$\left(\hat{x}_1, \hat{x}_2\right) = \arg \max_{(x_1, x_2) \in \mathbf{C}^2} \Pr\left(r_1, r_2 \middle| x_1, x_2\right) \tag{17}$$

Furthermore, since the noise n_1 and n_2 at time t and time $t+T$, respectively, are assumed to be mutually independent, we can alternatively express (17) as

$$\left(\hat{x}_1, \hat{x}_2\right) = \arg \max_{(x_1, x_2) \in \mathbf{C}^2} \Pr\left(r_1 \middle| x_1, x_2\right) \Pr\left(r_2 \middle| x_1, x_2\right) \tag{18}$$

Recall from (13) and (14) that r_1 and r_2 are two independent Gaussian random variables with distributions $r_1 \sim N\left(h_1x_1 + h_2x_2,\ \sigma^2\right)$ and $r_2 \sim N\left(-h_1x_2^* + h_2x_1^*,\ \sigma^2\right)$. Substituting this into (18), we then obtain a ML decoding criterion:

$$\begin{aligned}\left(\hat{x}_1, \hat{x}_2\right) &= \arg \min_{(x_1,x_2)\in \mathbf{C}^2} \left|r_1 - h_1x_1 - h_2x_2\right|^2 + \left|r_2 + h_1x_2^* - h_2x_1^*\right|^2 \\ &= \arg \min_{(x_1,x_2)\in \mathbf{C}^2} d^2\left(r_1, h_1x_1 + h_2x_2\right) + d^2\left(r_2, -h_1x_2^* + h_2x_1^*\right)\end{aligned} \tag{19}$$

where $d^2\left(s_1, s_2\right)$ denotes the Euclidean distance between s_1 and s_2. The ML decoder is, therefore, equivalent to choosing a pair of data symbols $\left(\hat{x}_1,\ \hat{x}_2\right)$ to minimize the distance metric, as indicated in (19). By replacing (13) and (14) into (19), the ML decoding criterion can be further rewritten as a meaningful expression as follows:

$$\left(\hat{x}_1, \hat{x}_2\right) = \arg \min_{(x_1,x_2)\in \mathbf{C}^2} \left(\left|h_1\right|^2 + \left|h_2\right|^2 - 1\right)\left(\left|x_1\right|^2 + \left|x_2\right|^2\right) + d^2\left(\tilde{x}_1, x_1\right) + d^2\left(\tilde{x}_2, x_2\right) \tag{20}$$

where $\tilde{x}_1$ and $\tilde{x}_2$ are two decision statistics obtained by combining the received signals r_1 and r_2 with channel state information h_1 and h_2, given by

$$\begin{aligned}\tilde{x}_1 &= h_1^* r_1 + h_2 r_2^* \\ \tilde{x}_2 &= h_2^* r_1 + h_1 r_2^*\end{aligned} \tag{21}$$

By taking r_1 and r_2 from equation (13) and (14), respectively, into (21), the decision statistics is given by

$$\begin{aligned}\tilde{x}_1 &= \left(\left|h_1\right|^2 + \left|h_2\right|^2\right)x_1 + h_1^* n_1 + h_2 n_2^* \\ \tilde{x}_2 &= \left(\left|h_1\right|^2 + \left|h_2\right|^2\right)x_2 - h_1 n_2^* + h_2^* n_1\end{aligned} \tag{22}$$

It is observed that for a given channel realization h_1 and h_2, the decision statistics $\tilde{x}_i$ in (22) is only a function of x_i, for $i = 1,\ 2$. Consequently, the ML decoding criterion in (20) can be divided into two independent decoding criteria for x_1 and x_2; that is, we have

$$\hat{x}_1 = \arg\min_{x_1\in \mathbf{C}} \left(\left|h_1\right|^2 + \left|h_2\right|^2 - 1\right)\left|x_1\right|^2 + d^2\left(\tilde{x}_1, x_1\right) \tag{23}$$

and

$$\hat{x}_2 = \arg\min_{x_2\in \mathbf{C}} \left(\left|h_1\right|^2 + \left|h_2\right|^2 - 1\right)\left|x_2\right|^2 + d^2\left(\tilde{x}_2, x_2\right) \tag{24}$$

Particularly, if a constant envelope modulation scheme such as M-phase-shift-keying (M-PSK) is adopted, the term $\left(\left|h_1\right|^2 + \left|h_2\right|^2 - 1\right)\left|x_i\right|^2$, for $i = 1,\ 2$, remains unchanged for all possible signal points with a fixed channel fading coefficients h_1 and h_2. Under this circumstance, the decision rules of (23) and (24) can be further simplified as

$$\hat{x}_1 = \arg\min_{x_1\in \mathbf{C}}\ d^2\left(\tilde{x}_1, x_1\right); \qquad \hat{x}_2 = \arg\min_{x_2\in \mathbf{C}}\ d^2\left(\tilde{x}_2, x_2\right) \tag{25}$$

From (25), for the case of constant envelope modulation, the decoding algorithm is just a linear decoder with extremely low complexity to achieve diversity gains. On the other hand, when non-constant envelope modulation, e.g., quadrature-amplitude-modulation (QAM) is adoped, the term $\left(|h_1|^2+|h_2|^2-1\right)|x_i|^2$, for $i=1,\,2$, may become different for various constellation points and cannot be excluded from the decoding metric. Therefore, we should follow the decoding rules as shown in (23) and (24) to achieve the ML decoding.

2.2.3 Alamouti's scheme with multiple receive antennas

We now extend the Alamouti's scheme to an MIMO communication system with n_R multiple receive antennas. Let us denote r_1^j and r_2^j as the received signals at the jth receive antenna at the time slot t and $t+T$, respectively. According to (13) and (14), it follows

$$\begin{aligned} r_1^j &= h_{j,1}x_1 + h_{j,2}x_2 + n_1^j \\ r_2^j &= -h_{j,1}x_2^* + h_{j,2}x_1^* + n_2^j \end{aligned} \tag{26}$$

where $h_{j,i}$, for $i=1,\,2$ and $j=1,\cdots,n_R$, is the channel fading gain from the transmit antenna i to the receive antenna j, and n_1^j and n_2^j are assumed to be spatially and temporally white Gaussian noises for the receive antenna j at time t and $t+T$, respectively. Similar to the derivation in the case of single receive antenna, the ML decoding criterion with multiple receive antennas now can be formulated as below:

$$\begin{aligned} \left(\hat{x}_1,\hat{x}_2\right) &= \arg\min_{(x_1,x_2)\in C^2}\sum_{j=1}^{n_R}\left|r_1^j - h_{j,1}x_1 - h_{j,2}x_2\right|^2 + \left|r_2^j + h_{j,1}x_2^* - h_{j,2}x_1^*\right|^2 \\ &= \arg\min_{(x_1,x_2)\in C^2}\sum_{j=1}^{n_R} d^2\left(r_1^j, h_{j,1}x_1 + h_{j,2}x_2\right) + d^2\left(r_2^j, -h_{j,1}x_2^* + h_{j,2}x_1^*\right) \end{aligned} \tag{27}$$

We then define two decision statistics by combining the received signals at each receive antenna with the corresponding channel link gains, as follows:

$$\begin{aligned} \tilde{x}_1^j &= h_{j,1}^* r_1^j + h_{j,2}\left(r_2^j\right)^* \\ \tilde{x}_{22}^j &= h_{j,2}^* r_1^j - h_{j,1}\left(r_2^j\right)^* \end{aligned} \tag{28}$$

Note that by replacing r_1^j and r_2^j, given in (26), into (28), the decision statistics can be explicitly written as

$$\begin{aligned} \tilde{x}_1^j &= \left(\left|h_{j,1}\right|^2 + \left|h_{j,2}\right|^2\right)x_1 + h_{j,1}^* n_1^j + h_{j,2}\left(n_2^j\right)^* \\ \tilde{x}_2^j &= \left(\left|h_{j,1}\right|^2 + \left|h_{j,2}\right|^2\right)x_2 - h_{j,1}\left(n_2^j\right)^* + h_{j,2}^* n_1^j \end{aligned} \tag{29}$$

where $G_{eff} = \left|h_{j,1}\right|^2 + \left|h_{j,2}\right|^2$ is the effective channel fading gain, and it is shown that the Alamouti's STC scheme can therefore extract a diversity order of two at each receiving branch, even in the absence of channel state information at the transmitter side. Following the derivation in (19) and (20), the ML decoding rules, under the case of n_R receive antennas, for the two data symbols x_1 and x_2 can be represented by

$$\hat{x}_1 = \arg\min_{x_1 \in \mathbf{C}} \sum_{j=1}^{n_R} \left(\left|h_{j,1}\right|^2 + \left|h_{j,2}\right|^2 - 1 \right) \left|x_1\right|^2 + d^2(\tilde{x}_1^j, x_1) \tag{30}$$

and

$$\hat{x}_2 = \arg\min_{x_2 \in \mathbf{C}} \sum_{j=1}^{n_R} \left(\left|h_{j,1}\right|^2 + \left|h_{j,2}\right|^2 - 1 \right) \left|x_2\right|^2 + d^2(\tilde{x}_2^j, x_2) \tag{31}$$

In particular, for constant envelope modulation schemes whose constellation points possess equal energy, the ML decoding can be reduced to finding a data symbol $\hat{x}_i$, for $i = 1,\ 2$, to minimize the summation of Euclidean distance $d^2(\tilde{x}_1^j, x_1)$ over all receive antennas, in the following:

$$\hat{x}_1 = \arg\min_{x_1 \in \mathbf{C}} \sum_{j=1}^{n_R} d^2\left(\tilde{x}_1^j, x_1\right) \tag{32}$$

and

$$\hat{x}_2 = \arg\min_{x_2 \in \mathbf{C}} \sum_{j=1}^{n_R} d^2\left(\tilde{x}_2^j, x_2\right) \tag{33}$$

2.2.4 BER performance of Alamouti's scheme

The bit error rate (BER) performance of the Alamouti's transmit diversity scheme is simulated and compared with the MRC receive diversity scheme in the following. At the beginning, it is assumed that a flat fading channel is used, and the fading gains from each transmit antenna to each receive antenna are mutually independent. Furthermore, we assume that the total transmission power from the two transmit antennas for the Alamouti's scheme is the same as that for the MRC receive diversity scheme.

Fig. 6 compares the BER performance between the Alamouti's with one or two receive antennas and the MRC receive diversity with two or four receive antennas. The Alamouti's scheme with two transmit antennas and one receive antenna has the same diversity order as the MRC receive diversity scheme with two branches. In other words, the slopes of these two BER performance curves are identical. However, the Alamouti's scheme has 3dB loss in terms of $\mathrm{E_b/N_0}$. This is due to the fact that for fair comparisons, the total transmission power is fixed and the energy radiated from each transmit antenna in the Alamouti's scheme is a half of that from a single transmit antenna in the MRC receive diversity scheme. Similarly, the Alamouti's scheme with two receive antennas can introduce the same diversity order as the MRC receiver diversity scheme with four branches, while there is still 3dB loss in BER performance. In general, the Alamouti's scheme with two transmit antennas and n_R receive antennas can provide a diversity order of $2 \times n_R$, which is the same as the case that the MRC scheme uses $2n_R$ receive antennas.

In Fig. 7, it is shown that the BER performance of the Alamouti's scheme with quadrature phase-shift keying (QPSK) modulation over flat fading channels. It is obvious that the more number of receive antennas it uses, the higher diversity order it can achieves.

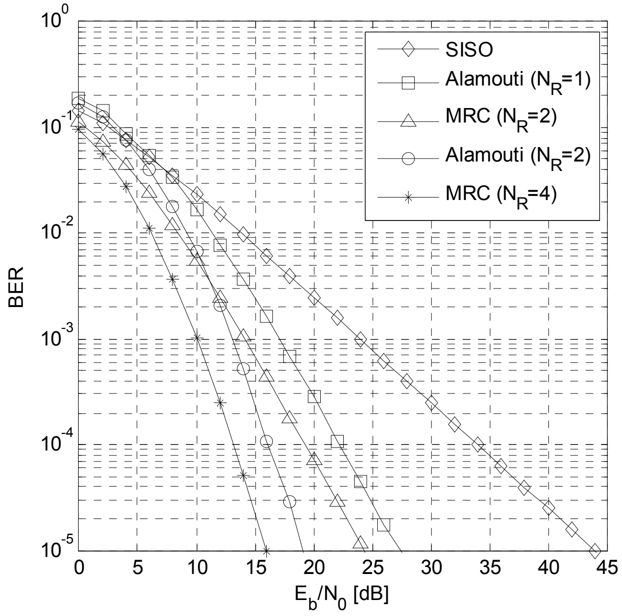


Fig. 6. Comparison of BER performance between Alamouti's and MRC schemes with binary phase-shift keying (BPSK) modulation

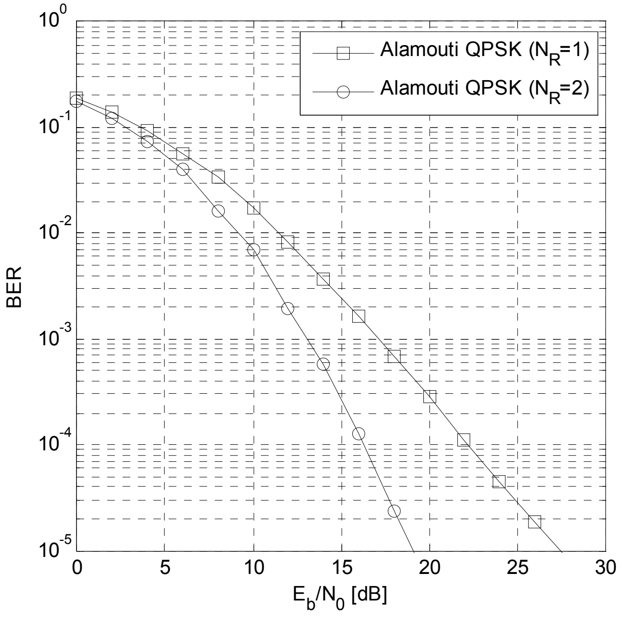


Fig. 7. BER performance of Alamouti's scheme using QPSK modulation

2.2.5 Generalized space-time block codes

As we discussed, the Alamouti's scheme shows a very elegant way to achieve full diversity gains, i.e., a diversity order of two, with a low-complexity linear decoding algorithm by utilizing two transmit antennas. The key feature of the Alamouti's scheme is the orthogonal property of the encoding matrix in (1), i.e., the sequences generated by the two transmit antennas are independent of each other. In (Tarokh et al., 1999), Tarokh generalizes this idea to any arbitrary number of transmit antennas by applying the orthogonal design theory, and proposes a series of STBCs which can fulfill transmit diversity specified by the number of transmit antennas n_T. Meanwhile, these STBCs also enable a very simple maximum-likelihood decoding algorithm, based only on a linear processing of the received signals at different time slots.

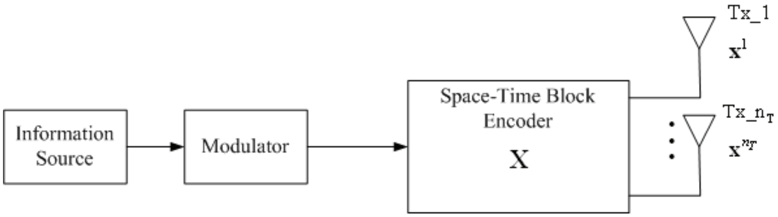


Fig. 8. Encoder structure of STBCs

The encoder structure for generalized STBCs is presented in Fig. 8. In general, a STBC can be defined via an $n_T \times p$ transmission matrix $\mathbf{X}$, where n_T represents the number of transmit antennas and p is the time duration for transmitting each block of space-time coded symbols. Consider a M-ary modulation scheme, where we define $m = \log_2 M$ as the number of information bits required for each constellation point mapping. At each encoding operation, a block of km information bits are mapped onto k modulated data symbols x_i, for $i = 1,\ldots,k$. Subsequently, these k modulated symbols are encoded by the $n_T \times p$ space-time encoder $\mathbf{X}$ to generate n_T parallel signal sequences of length p which are to be transmitted over n_T transmit antennas simultaneously within p time slots. The code rate of a STBC is defined as the ratio of the number of symbols taken by the space-time encoder as its input to the number of space-time coded symbols transmitted from each antenna. Since p time slots are required for transmitting k information-bearing data symbols, the code rate is given by

$$R_c \triangleq \frac{k}{p} \tag{34}$$

Therefore, the spectral efficiency for the STBC is calculated by

$$\eta \triangleq \frac{r_b}{B} = \frac{(r_s m) R_c}{r_s} = \frac{km}{p} \ (\text{bits/sec/Hz}) \tag{35}$$

where r_b and r_s are the bit and symbol rate of a space-time coded symbol, respectively, and B represents the total bandwidth. For simplicity of notations, we usually denote a STBC with n_T transmit antennas as $\mathbf{X}_{n_T}$. Based on the orthogonal designs in (Tarokh et al., 1999), to obtain full diversity gains, i.e., diversity order is equal to n_T, the space-time encoding matrix should preserve the orthogonal structure; that is, we have

$$\mathbf{X}_{n_T} \cdot \mathbf{X}_{n_T}^{\dagger} = c\left(|x_1|^2 + |x_2|^2 + \cdots + |x_k|^2\right)\mathbf{I}_{n_T} \tag{36}$$

where c is constant, $(\bullet)^{\dagger}$ takes the Hermitian operation, and $\mathbf{I}_{n_T}$ is an $n_T \times n_T$ identity matrix, the entries in $\mathbf{X}_{n_T}$ take the values of modulated symbols x_i, their conjugate x_i^*, or their combination. The orthogonal structure allows the receiver to decouple the signals transmitted from different antennas by using a simple linear decoder derived based on the ML decoding metric. Tarokh et al. (Tarokh et al., 1999) discovered that the code rate of a STBC with full diversity must be less than or equal to one, i.e., $R_c \le 1$. In other words, the STBCs cannot be used to increase bandwidth efficiency, but provide diversity gains. It is noted that the full code rate, $R_c = 1$, requires no additional bandwidth expansion, while the code rate $R_c \le 1$ requires a bandwidth expansion by a factor of $1/R_c$.

Based on modulation types, STBCs can be classified into two categories: real signaling or complex signaling. For a special case of $p = n_T$, it is evident from (Tarokh et al., 1999) that for an arbitrary real constellation signaling, e.g., M-amplitude shift keying (M-ASK), STBCs with an $n_T \times n_T$ square encoding matrices $\mathbf{X}_{n_T}$ exist if and only if the number of transmit antennas n_T is equal to two, four, or eight. Moreover, these code matrices can not only achieve the full code rate $R_c = 1$ but also provide the full diversity gains with a diversity order of n_T. However, it is desirable to have code matrices with the full diversity gains and the full code rate for any number of transmit antennas. It has been proved that for n_T transmit antennas, the minimum required value for the transmission periods p to achieve the full diversity n_T and the full code rate $R_c = 1$ must satisfy the following condition:

$$\min_{\{(c,d)|0\le c,\ 0\le d\le 4,\ \text{and}\ 8c+2^d\ge n_T\}} 2^{4c+d} \tag{37}$$

n_T	p	$\mathbf{X}_{n_T}$
2	2	$\mathbf{X}_2 = \begin{bmatrix} x_1 & -x_2 \\ x_2 & x_1 \end{bmatrix}$
4	4	$\mathbf{X}_4 = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 \\ x_2 & x_1 & x_4 & -x_3 \\ x_3 & -x_4 & x_1 & x_2 \\ x_4 & x_3 & x_2 & x_1 \end{bmatrix}$
8	8	$\mathbf{X}_8 = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 & -x_5 & -x_6 & -x_7 & -x_8 \\ x_2 & x_1 & -x_4 & x_3 & -x_6 & x_5 & x_8 & -x_7 \\ x_3 & x_4 & x_1 & -x_2 & -x_7 & -x_8 & x_5 & x_6 \\ x_4 & -x_3 & x_2 & x_1 & -x_8 & x_7 & -x_6 & x_5 \\ x_5 & x_6 & x_7 & x_8 & x_1 & -x_2 & -x_3 & -x_4 \\ x_6 & -x_5 & x_8 & -x_7 & x_2 & x_1 & x_4 & -x_3 \\ x_7 & -x_8 & -x_5 & x_6 & x_3 & -x_4 & x_1 & x_2 \\ x_8 & x_7 & -x_6 & -x_5 & x_4 & x_3 & -x_2 & x_1 \end{bmatrix}$

Table 1. Square code matrices with full diversity gains and full code rate for $n_T = 2,\ 4,\ 8$

Accordingly, the minimum values of p for a specific value of $n_T \le 8$, and the associated STBC matrices $\mathbf{X}_{n_T}$ for real signaling are provided as follows, where the square transmission matrices $\mathbf{X}_2$, $\mathbf{X}_4$, and $\mathbf{X}_8$ are listed in Table 1, and the non-square transmission matrices $\mathbf{X}_3$, $\mathbf{X}_5$, $\mathbf{X}_6$ and $\mathbf{X}_7$ are listed in Table 2.

n_T	p	$\mathbf{X}_{n_T}$
3	4	$\mathbf{X}_3 = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 \\ x_2 & x_1 & x_4 & -x_3 \\ x_3 & -x_4 & x_1 & x_2 \end{bmatrix}$
5	8	$\mathbf{X}_5 = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 & -x_5 & -x_6 & -x_7 & -x_8 \\ x_2 & x_1 & -x_4 & x_3 & -x_6 & x_5 & x_8 & -x_7 \\ x_3 & x_4 & x_1 & -x_2 & -x_7 & -x_8 & x_5 & x_6 \\ x_4 & -x_3 & x_2 & x_1 & -x_8 & x_7 & -x_6 & x_5 \\ x_5 & x_6 & x_7 & x_8 & x_1 & -x_2 & -x_3 & -x_4 \end{bmatrix}$
6	8	$\mathbf{X}_6 = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 & -x_5 & -x_6 & -x_7 & -x_8 \\ x_2 & x_1 & -x_4 & x_3 & -x_6 & x_5 & x_8 & -x_7 \\ x_3 & x_4 & x_1 & -x_2 & -x_7 & -x_8 & x_5 & x_6 \\ x_4 & -x_3 & x_2 & x_1 & -x_8 & x_7 & -x_6 & x_5 \\ x_5 & x_6 & x_7 & x_8 & x_1 & -x_2 & -x_3 & -x_4 \\ x_6 & -x_5 & x_8 & -x_7 & x_2 & x_1 & x_4 & -x_3 \end{bmatrix}$
7	8	$\mathbf{X}_7 = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 & -x_5 & -x_6 & -x_7 & -x_8 \\ x_2 & x_1 & -x_4 & -x_3 & -x_6 & x_5 & x_8 & -x_7 \\ x_3 & x_4 & x_1 & -x_2 & -x_7 & -x_8 & x_5 & x_6 \\ x_4 & -x_3 & x_2 & x_1 & -x_8 & x_7 & -x_6 & x_5 \\ x_5 & x_6 & x_7 & x_8 & x_1 & -x_2 & -x_3 & -x_4 \\ x_6 & -x_5 & x_8 & -x_7 & x_2 & x_1 & x_4 & -x_3 \\ x_7 & -x_8 & -x_5 & x_6 & x_3 & -x_4 & x_1 & x_2 \end{bmatrix}$

Table 2. Non-square code matrices with full diversity gains and full code rate for n_T = 3, 5, 6, 7

The other type of STBCs belongs to complex constellation signaling, and just as the case for the real constellation signaling, these complex STBCs also abide by the orthogonal design constraint in (36). In particular, Alamouti's scheme can be regarded as a complex STBC for two transmit antennas; that is, the code matrix can be expressed as

$$\mathbf{X}_2^{\mathrm{C}} = \begin{bmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{bmatrix} \tag{38}$$

where we use $\mathbf{X}_{n_T}^{C}$ to denote a complex STBC for n_T transmit antennas in order to discriminate between real and complex matrices. It is noted that the Alamouti's scheme can provide a diversity order of two and the full code rate. As compared with those real STBCs, it is much more desirable to invent complex STBCs since complex constellation schemes usually exhibit higher bandwidth efficiency. In addition, one might wonder if there exist other complex STBC matrices for $n_T > 2$. Unfortunately, it has been shown that Alamouti's scheme is the only complex STBC with an $n_T \times n_T$ square code matrix to simultaneously achieve the full diversity gains and the full code rate. For the case of $n_T > 2$, we might intend to construct complex orthogonal matrices $\mathbf{X}_{n_T}^{C}$ that can achieve full diversity gains, but with high code rate R_c and minimum decoding latency p . We summarize below that for any number of transmit antennas, there exist complex STBCs with code rate of $R_c = 1/2$. For example, for the cases of three or four transmit antennas, the code matrices are given by

$$\mathbf{X}_3^C = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 & x_1^* & -x_2^* & -x_3^* & -x_4^* \\ x_2 & x_1 & x_4 & -x_3 & x_2^* & x_1^* & x_4^* & -x_3^* \\ x_3 & -x_4 & x_1 & x_2 & x_3^* & -x_4^* & x_1^* & x_2^* \end{bmatrix}, \; R_c = 1/2 \tag{39}$$

$$\mathbf{X}_4^C = \begin{bmatrix} x_1 & -x_2 & -x_3 & -x_4 & x_1^* & -x_2^* & -x_3^* & -x_4^* \\ x_2 & x_1 & x_4 & -x_3 & x_2^* & x_1^* & x_4^* & -x_3^* \\ x_3 & -x_4 & x_1 & x_2 & x_3^* & -x_4^* & x_1^* & x_2^* \\ x_4 & x_3 & -x_2 & x_1 & x_4^* & x_3^* & -x_2^* & x_1^* \end{bmatrix}, \; R_c = 1/2 \tag{40}$$

Still, for the case of $n_T > 2$, we can acquire other higher code-rate complex code matrices at the expense of complicated linear encoding and decoding processing. For example, the following two matrices $\mathbf{X}_3^C$ and $\mathbf{X}_4^C$ are STBCs with a code rate of $R_c = 3/4$:

$$\mathbf{X}_3^{C,3/4} = \begin{bmatrix} x_1 & -x_2^* & \dfrac{x_3^*}{\sqrt{2}} & \dfrac{x_3^*}{\sqrt{2}} \\ x_2 & x_1^* & \dfrac{x_3^*}{\sqrt{2}} & \dfrac{-x_3^*}{\sqrt{2}} \\ \dfrac{x_3}{\sqrt{2}} & \dfrac{x_3}{\sqrt{2}} & \dfrac{\left(-x_1 - x_1^* + x_2 - x_2^*\right)}{2} & \dfrac{\left(x_2 + x_2^* + x_1 - x_1^*\right)}{2} \end{bmatrix}, \; R_c = 3/4 \tag{41}$$

$$\mathbf{X}_4^{C,3/4} = \begin{bmatrix} x_1 & -x_2 & \dfrac{x_3^*}{\sqrt{2}} & \dfrac{x_3^*}{\sqrt{2}} \\ x_2 & x_1 & \dfrac{x_3^*}{\sqrt{2}} & \dfrac{-x_3^*}{\sqrt{2}} \\ \dfrac{x_3}{\sqrt{2}} & \dfrac{x_3}{\sqrt{2}} & \dfrac{\left(-x_1 - x_1^* + x_2 - x_2^*\right)}{2} & \dfrac{\left(x_2 + x_2^* + x_1 - x_1^*\right)}{2} \\ \dfrac{x_3}{\sqrt{2}} & \dfrac{-x_3}{\sqrt{2}} & \dfrac{\left(-x_2 - x_2^* + x_1 - x_1^*\right)}{2} & \dfrac{-\left(x_1 + x_1^* + x_2 - x_2^*\right)}{2} \end{bmatrix}, \; R_c = 3/4 \tag{42}$$

2.3 Maximum ratio transmission techniques

The multipath fading is considered as a detrimental effect to degrade the performance of wireless communications. The most common and simplest way to mitigate the multipath fading is to adopt antenna diversity techniques. As we introduced in the previous section, with multiple receive antennas, the MRC scheme can attain receive diversity gains and maximize the output SNR, while with multiple transmit antennas, STC schemes are proposed to obtain the MRC-like transmit diversity gains. For example, Alamouti's STC can obtain a diversity order of two by encoding a pair of two symbols to transmit over two transmit antennas and two contiguous time slots. However, these transmit diversity techniques are designed with the object of providing diversity gains, other than to maximize the post-output SNR at the receiver side, which is usually taken as one of the most important performance figure to minimize the BER performance. Furthermore, more and more wireless communication systems are now equipped with multiple transmit and receive antennas, and therefore it is desired to simultaneously obtain transmit and receive diversity gains so as to combat the severe fading effects.

In this section, we introduce a MRT scheme to fulfill the above two challenges, namely achieving transmit and receive diversity gains and maximizing the post-output SNR (Lo, 1999). Finally, numerical results are presented.

2.3.1 MRT systems and schemes

The developed framework of MRT schemes can be regarded as the generalization of MRC for the systems with multiple transmit-and-receive antenna pairs. Figure 9 presents the MRT system model for which n_T transmit antennas and n_R receive antennas are equipped. Accordingly, there are $n_T \times n_R$ channel links between the transmitter and the receiver. We assume channel coefficients are statistically independent, as follows:

$$\mathbf{H} = \begin{bmatrix} h_{11} & \cdots & h_{1n_T} \\ \vdots & \ddots & \vdots \\ h_{n_R 1} & \cdots & h_{n_R n_T} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{n_R} \end{bmatrix} \tag{43}$$

where the entry $h_{n_R n_T}$ denotes the flat fading channel gain from the $n_T th$ transmit antenna to the $n_R th$ receive antenna. In general, it is essential to acquire the channel state information for successful implementation of a multiple-input multiple-output communication system. The channel estimation issue is beyond the scope of this chapter, and it is assumed that channel state information of (43) is perfectly known at the both transmitter and receiver sides.

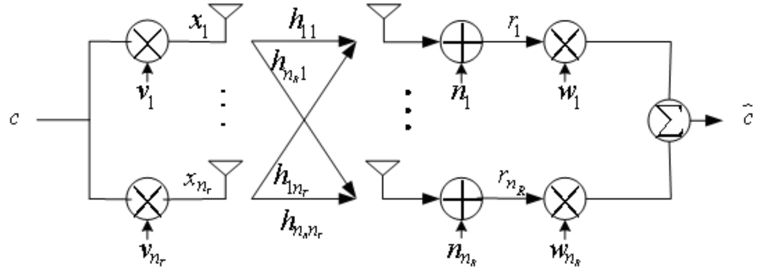


Fig. 9. System model of MRT

The MRT system is a diversity-achieving technique. For each time slot, only one data symbol, denoted by s, is transmitted by multiplying with antenna-specific weighting coefficients v_m, $m = 1,\ldots,n_T$. The average symbol power is assumed to be $E\left[|s|^2\right] = E_s$. Denoting the weighting coefficients vector as $\mathbf{v} = \left[v_1,\ldots,v_{n_T}\right]^T$, the signals transmitted by antennas can be expressed by

$$\mathbf{x} = \left[x_1,\ldots,x_{n_T}\right]^T = s\mathbf{v} \tag{44}$$

From (43) and (44), the received signals is given by

$$\mathbf{r} = \mathbf{Hx} + \mathbf{n} \tag{45}$$

where $\mathbf{n} = \left[n_1,\ldots,n_{n_R}\right]^T$ is the additive white Gaussian noise vector. Subsequently, a weighting vector $\mathbf{w} = \left[w_1,\ldots,w_{n_R}\right]^T$ is applied for combining the received signals at the receive antennas, followed by symbol decision. In the MRT system, we construct the weighting factor $\mathbf{v}$ from the channel matrix $\mathbf{H}$ through a linear transformation as follows:

$$\mathbf{v} = \frac{1}{a}(\mathbf{gH})^\dagger \tag{46}$$

where $\mathbf{g} = \begin{bmatrix} g_1 & \cdots & g_{n_R} \end{bmatrix}$ is a parameter to be determined. The transmitted signal can be alternatively written as

$$\mathbf{x} = \frac{1}{a}(\mathbf{gH})^\dagger s \tag{47}$$

where a is a normalization factor such that the transmit power is equal to one, which can be represented as

$$a = |\mathbf{gH}| = \left(\sum_{p=1}^{n_R}\sum_{q=1}^{n_R} g_p g_q^* \sum_{k=1}^{n_T} h_{pk} h_{gk}^*\right)^{1/2} \tag{48}$$

Therefore, from (47), the received signal vector is given by

$$\mathbf{r} = \frac{1}{a}\mathbf{H}(\mathbf{gH})^\dagger s + \mathbf{n} \tag{49}$$

In order to make symbol decision, the weighting vector $\mathbf{w}$ is applied to combine the received signal vector $\mathbf{x}$ for signal reception. If $\mathbf{w}^T$ is equal to $\mathbf{g}$, the output is given by

$$\tilde{s} = \mathbf{w}^T\mathbf{r} = \mathbf{g}\frac{1}{a}\mathbf{H}(\mathbf{gH})^\dagger s + \mathbf{gn} = as + \mathbf{gn} \tag{50}$$

The post-output SNR after combining is thus calculated as

$$\gamma = \frac{a^2}{\mathbf{gg}^\dagger}\gamma_0 = \frac{a^2\gamma_0}{\sum_{p=1}^{n_R}|g_p|^2} \tag{51}$$

where $\gamma_0 = E_s / \sigma_n^2$ denotes the average SNR for a single transmit antenna. It is observed from (51) that the post-output SNR only depends on the value of $\mathbf{g}$. Thus, to maximize the post-output SNR is equivalent to choosing an appropriate value of $\mathbf{g}$. By assuming all channel links between the transmitter and receiver are statistically identical, each of the transmitter antenna has to be allocated with the same weighting power; that is, we have $|g_1| = |g_2| = \cdots = |g_{n_R}|$. Without loss of generality, we can set $|g_i| = 1$, for $i = 1, \cdots, n_R$, for simplicity. Accordingly, the post-output SNR in (51) is simplified as

$$\gamma = \frac{a^2}{N} \gamma_0 \tag{52}$$

By applying the Cauchy-Schwarz inequality, we can further get that a^2 is maximized, i.e., maximizing the post-output SNR, if and only if we set

$$\left(g_p g_q^* \right)^* = \frac{\sum_{k=1}^{n_T} h_{pk} h_{gk}^*}{\left| \sum_{k=1}^{n_T} h_{pk} h_{gk}^* \right|} \tag{53}$$

It is noted that the denominator term in (53) is due to the fact of $|g_i| = 1$ and $\left| g_p g_q^* \right| = 1$. In accordance with (52) and (53), the maximum post-output SNR is given by

$$\begin{aligned} \gamma_{\max} &= \frac{\gamma_0}{N} \sum_{p=1}^{n_R} \sum_{q=1}^{n_R} \left| \sum_{k=1}^{n_T} h_{pk} h_{gk}^* \right| \\ &= \frac{\gamma_0}{N} \sum_{p=1}^{n_R} \sum_{q=1}^{n_R} \left| \mathbf{h}_p \mathbf{h}_q^{\dagger} \right| \end{aligned} \tag{54}$$

On the one hand, if $\mathbf{h}_p$ and $\mathbf{h}_q$ are mutually orthogonal, i.e., $\mathbf{h}_p \mathbf{h}_q^H = 0$, $\gamma_{\max}$ takes the minimum value, given by

$$\gamma_{\max} = \frac{\gamma_0}{N} \sum_{p=1}^{n_R} \sum_{k=1}^{n_T} \left| h_{pk} \right|^2 \tag{55}$$

On the other hand, if $\mathbf{h}_p$ and $\mathbf{h}_q$ are fully correlated, i.e., $\mathbf{h}_p \mathbf{h}_q^H = \left| \mathbf{h}_q \right|^2$, $\gamma_{\max}$ takes the maximum value, given by

$$\gamma_{\max} = \frac{\gamma_0}{N} \sum_{p=1}^{n_R} \sum_{q=1}^{n_R} \sum_{k=1}^{n_T} \left| h_{pk} \right|^2 \tag{56}$$

Form (55) and (56), the average post-output SNR is therefore bounded by

$$M \bar{\gamma}_0 \le \bar{\gamma}_{\max} \le N M \bar{\gamma}_0 \tag{57}$$

where $\bar{\gamma}_0 = \gamma_0 E\left[\left| h_{pk} \right|^2 \right]$ is defined as the average SNR at each branch. Also, for an MRT system with $n_T \times n_R$ antennas, it is expected from (55) that the diversity order is equal to $n_T \times n_R$.

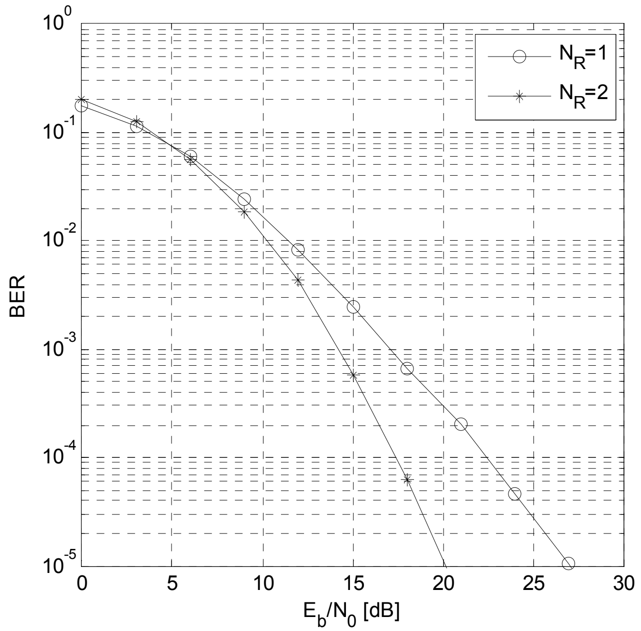


Fig. 10. Comparison of MRT schemes with $N_T = 2$ and QPSK modulation

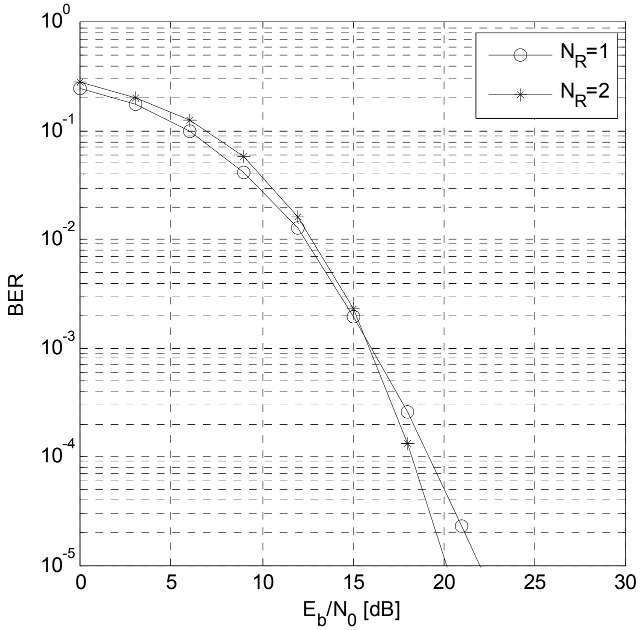


Fig. 11. Comparison of MRT schemes with $N_T = 4$ and QPSK modulation

2.3.2 BER performance of MRT scheme

The BER performance of the MRT scheme is simulated as follows. At the beginning, we assume that the complex Gaussian fading gains among transmit-receive pairs are mutually independent. The total channel power from all the transmit antennas to a receive antenna is normalized to one. The QPSK modulation is adopted in the simulation. Moreover, it is assumed that channel fading gains are perfectly available to both transmitter and receiver sides. Fig. 10 and Fig. 11 demonstrate the BER performance curves for an MRT system with two transmit antennas and two and four receive antennas, respectively. It is observed from these two figures that the BER performance can be improved as the number of transmit and receive antennas increases, and we can gain much diversity gains by increasing the number of antennas.

3. Spatial multiplexing techniques

Recent research resutls from information theory aspects have disclosed that rich scattering wireless channels can provide enormous capacity which increases proportional to the minimum number of transmit and receive antennas (Foschini, 1996). Spatial multiplexing based MIMO (SM-MIMO) systems, in which multiple data streams are simultaneously transmitted through multiple antennas, is a new attractive technique to realize high data rate transmission, while signal detection at the receiver is a difficult task for SM-MIMO systems. Although the well-known ML detector can be applied for the SM-MIMO systems achieve the best performance, its complexity increases exponentially as the number of data stream and the number of transmit antennas increase.

Foschini proposed a layered space-time (LST) or Bell laboratories layed space-time (BLAST) architecture (Foschini & Gans, 1998), allowing for processing multi-dimensional received signals in spatial domain. At the transmitter, multiple data streams are first encoded and then distributed over multiple antennas. At the reciever, the received signals are separated and subsequently decoded by integrating interference supression or cancellation tehcniques with decoding algorithms, thus leading to much lower computation complexity, as compared with the ML decoding. Various BLAST architectures, depending on whether error control coding is adopted or not, have been investigated, e.g., horizontal BLAST (H-BLAST), diagonal BLAST (D-BLAST), and vertical BLAST (V-BLAST). The error control coding is applied for both the H-BLAST and D-BLAST architectures, while the V-BLAST does not consider the use of error control coding. Although the H-BLAST and D-BLAST architectures can gain better performance due to the coding gains in spatial domain, they suffer from a spectral efficiency loss and higher implementation complexity (Foschini & Gans, 1998). Therefore, the uncoded and simplified V-BLAST architecture, proposed in (Wolniansky et al., 1998), is considered as an effective solution to fullfill the potential spatial multiplexing gains. In this section, we focus on the V-BLAST architecture, followed by data detection techniques used to separate and detect the V-BLAST signals. Two essential V-BLAST receivers based on ZF and MMSE are introduced. Finally, we discuss the BER performance of the V-BLAST architecture with different data detection algorithms.

3.1 V-BLAST systems

The block diagram of the V-BLAST system is presented in Fig. 12, where there are n_T transmit and n_R receive antennas and $n_T \leq n_R$. A single data stream is divided into n_T substreams, each of which is then individually modulated into symbols and transmitted from its corresponding transmit antenna. Without loss of generality, it is assumed that the same constellation is applied for all substreams, and the transmission duration is a burst consisting of L symbols. We assume that the multiple-input multiple-output channel is flat fading and quasi-static over the duration of L symbols, and the channel matrix is denoted by $\mathbf{H}_{n_R \times n_T}$, whose $(i,j)th$ entry h_{ij} represents the complex channel gain from the jth transmit antenna to the ith receive antenna. We denote the data vector to be transmitted as

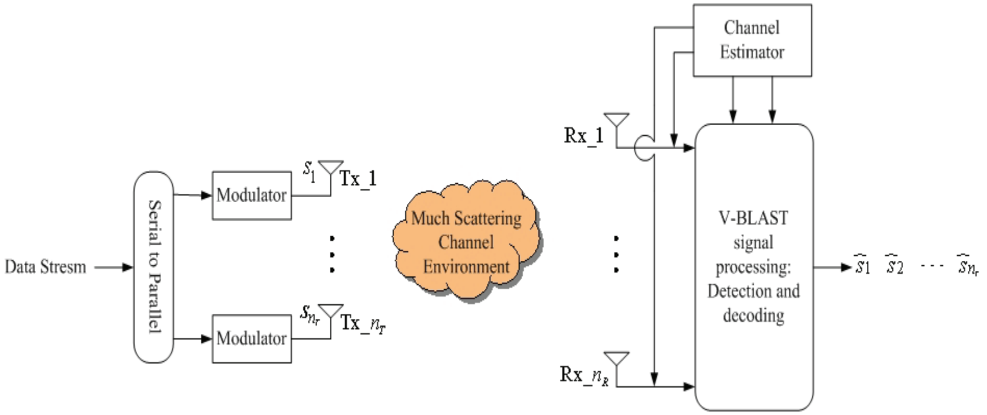

Channel Estimator
Rx_1
s_1
Tx_1
Data Stresm
Serial to Parallel
Modulator
Much Scattering Channel Environment
V-BLAST signal processing: Detection and decoding
s_{n_r}
Tx_n_T
Modulator
Rx_n_R
$\hat{s}1$ $\hat{s}2$ $\cdots$ $\hat{s}n_r$

Fig. 12. Block diagram of V-BLAST systems

$$\mathbf{s} = \left[s_1, \ldots, s_{n_T} \right]^T \tag{58}$$

where s_i is the modulated symbol at the ith transmit antenna. The received signals at the n_R receive antennas can thus be represented as a vector, as follows:

$$\mathbf{r} = \mathbf{Hs} + \mathbf{n} \tag{59}$$

where $\mathbf{n}$ is an independent and identical distributied (i.i.d.) white Gaussian noise vector with zero mean and a covariance matrix $\sigma_n^2 \mathbf{I}$.

3.2 V-BLAST data detection

The idea to perform data detection for this system is to incorporate conventional linear equalizers with nonlinear interference cancellation methods. Each substream is in trun detected, and the reminders are considered as interference signals. The equalizer is designed according to some specific criteria, e.g., ZF and MMSE, to null out the interference signals by linearly weighting the received signals with equalizer coefficients. In the following, we discuss the ZF-based V-BLAST receiver and the MMSE-based V-BLAST receiver.

3.2.1 ZF-based V-BLAST algorithm

First of all, the ZF nulling is performed by strictly choosing nulling vectors $\mathbf{w}_i$, for $i = 1, \ldots, M$, such that

$$\mathbf{w}_i^\dagger (\mathbf{H})_j = \delta_{ij}, \text{ for } i,j = 1,2,\cdots n_T \tag{60}$$

where $(\mathbf{H})_j$ denotes the jth column of $\mathbf{H}$, the notation $(\bullet)^\dagger$ takes the Hermitian, and δ is the Kronecker delta function. Then, the ith substream after equalization is given by

$$y_i = \mathbf{w}_i^\dagger \mathbf{r} \tag{61}$$

The ZF-based detection procedures to extract substreams for any arbitrary detection order are elaborated in the following. Let us set $j=1$, $\mathbf{r}_1 = \mathbf{r}$ and the ordering set for data detection as $\varsigma = \{k_1, k_2, \cdots, k_{n_T}\}$, where $k_j \in \{1, \ldots, n_T\}$.

Step 1. Use the nulling vector $\mathbf{w}_{k_j}$ to obtain the decision statistic for the $k_j th$ substream

$$y_{k_j} = \mathbf{w}_{k_j}^T \mathbf{r}_j \tag{62}$$

Step 2. Slice y_{k_j} to obtain the hard decision $\hat{s}_{k_j}$

$$\hat{s}_{k_j} = Q\left(y_{k_j}\right) \tag{63}$$

where $Q(\bullet)$ denotes the hard decision operation.

Step 3. Reconstruct and cancel out the currently detected substream from the received signal $\mathbf{r}_j$, resulting in a modified received signal vector $\mathbf{r}_{j+1}$

$$\mathbf{r}_{j+1} = \mathbf{r}_j - \hat{s}_{k_j} (\mathbf{H})_{k_j} \tag{64}$$

Update j to j+1 for the next iteration, and repeat the *Step1~Step3*, where the $k_j th$ ZF nulling vector is given by

$$\mathbf{w}_{k_j}^\dagger (\mathbf{H})_{k_i} = \begin{cases} 0, & \text{for } i \geq j \\ 1, & \text{for } i = j \end{cases} \tag{65}$$

Thus, if the inter-antenna interference is perfectly reconstructed and cancelled, the weighting vector $\mathbf{w}_{k_j}^\dagger$ is orthogonal to the subspace spanned by the interference-reduced vector $\mathbf{r}_{j+1}$. Accordingly, the solution to $\mathbf{w}_{k_j}^\dagger$ in (65) is the $k_j th$ row of $\widetilde{\mathbf{H}}_{j-1}^{+}$, where the notation $\widetilde{\mathbf{H}}_{j-1}$ denotes the matrix acquired by deleting columns $k_1, k_2, \cdots, k_{j-1}$ of $\mathbf{H}$ and $(\bullet)^+$ denotes the Moore-Penrose pseudoinverse (Abadir & Magnus, 2006). The post-detection SNR for the $k_j th$ substream of $\mathbf{s}$ is therefore given by

$$SNR_{k_j} = \frac{E_s}{\sigma_n^2 \left\|\mathbf{w}_{k_j}\right\|^2} \tag{66}$$

where $E_s = E\left[\left|s_{k_j}\right|^2\right]$.

3.2.2 MMSE-based V-BLAST algorithm

The MMSE is another well-known criterion for designing V-BLAST data detection. For the MMSE criterion, we intend to find the equalizer coefficients to minimize the mean squared error between the transmitted vector $\mathbf{s}$ and the equalized output $\mathbf{W}^\dagger \mathbf{r}$, as follows:

$$\mathbf{W}_{MMSE} = \arg\min_{\mathbf{W}} E\left\{(\mathbf{s} - \mathbf{W}^{\dagger}\mathbf{r})^2\right\} \tag{67}$$

The optimal MMSE equalizer $\mathbf{W}_{MMSE}$ is expressed as

$$\mathbf{W}_{MMSE} = \left(\mathbf{H}\mathbf{H}^{\dagger} + \sigma_n^2 \mathbf{I}_{n_R}\right)^{\text{-1}} \mathbf{H} \tag{68}$$

where σ_n^2 denotes the noise power and $\mathbf{I}_{n_R}$ represents an $n_R \times n_R$ identity matrix. Thus, the decision statistic for the ith substream is given by

$$y_i = (\mathbf{W}_{MMSE})_i^{\dagger}\mathbf{r} \tag{69}$$

where $(\mathbf{W}_{MMSE})_i$ denotes the ith column of the matrix $\mathbf{W}_{MMSE}$. Subsequently, the hard decision for the ith substream is given by

$$\hat{s}_i = Q(y_i) \tag{70}$$

To further improve the performance, one can incorporate the interference cancellation methodology, similar to the idea in subsection 3.2.1, into the MMSE equalizer. Concerning an arbitrary detection order, the interference suppression and cancellation procedure is the same as the *Step1~Step3* in subsection 3.2.1, but the MMSE equalizer is used instead of the nulling vector as follows. Define the MMSE equalizer at the jth iteration as $\mathbf{W}_{MMSE}^{j}$:

$$\mathbf{W}_{MMSE}^{j} = \left(\mathbf{H}_d^{j-1}\left(\mathbf{H}_d^{j-1}\right)^{\dagger} + \sigma_n^2 \mathbf{I}_{n_R}\right)^{\text{-1}} \mathbf{H}_d^{j-1} \tag{71}$$

where $\mathbf{H}_d^{j-1}$ represents the truncated channel matrix by deleting the columns $k_1, k_2, \cdots, k_{j-1}$ of $\mathbf{H}$. At the jth iteration, the weighting vector for detecting the $k_j th$ substream, $\left(\mathbf{W}_{MMSE}^{j}\right)_{k_j}$, is thus obtained from the $k_j th$ column of $\mathbf{W}_{MMSE}^{j}$.

3.2.3 Ordering V-BLAST algorithm

Since the ZF-based and MMSE-based V-BLAST data detection algorithms iterate between equalization and interference cancellation, the order for detecting the substreams of $\mathbf{s}$ becomes an important role to determine the overall performance (Foschini et al., 1999). In this subsection, we discuss an ordering scheme for the two V-BLAST detectoin algorithms.

Although the ZF or MMSE equalizer can null out or supress the residual inter-antenna interference, it will introudce the noise enhancement problem, leading to incorrect data decision, and the incorrect interference reconstruction will cause the error propagation problem. Assuming that all the substreams adopt the same constellation scheme, among all the remaining entries of $\mathbf{s}$ (not yet detected), the entry with the largest SNR, i.e., from (66), having the minimum norm power $\left\|\mathbf{w}_{k_j}\right\|^2$, is choosen at each iteration in the detection process. The iterative procedures for the ordering ZF-based or MMSE-based V-BLAST detection algorithms are described in the following.

```
Initialization Step:

Set $j=1$
Calculate $\mathbf{G}_1=\mathbf{H}^{+}$ (if ZF) or $\mathbf{G}_1=\left(\mathbf{H}\mathbf{H}^{\dagger}+\sigma_n^2\mathbf{I}_{n_R}\right)^{-1}\mathbf{H}$ (if MMSE)
Choose $k_1=\arg\min_i\left\|(\mathbf{G}_1)_i\right\|^2$
Recursion Step:

While $j\le n_T$
{
Interference supression & cancellation Part:
Calculate the weighting vector: $\mathbf{w}_{k_j}^{\dagger}=\left(\mathbf{G}_j\right)_{k_j}$ (if ZF) or $\mathbf{w}_{k_j}^{\dagger}=\left(\mathbf{G}_j\right)_{k_j}^{\dagger}$ (if MMSE)
Equalization: $y_{k_j}=\mathbf{w}_{k_j}^{\dagger}\mathbf{r}_j$
Slice: $\hat{s}_{k_j}=Q\left(y_{k_j}\right)$
Interference cancellation: $\mathbf{r}_{j+1}=\mathbf{r}_j-\hat{s}_{k_j}\left(\mathbf{H}\right)_{k_j}$
Ordering Part:
Calculate the equalizer matrix of the updated channel matrix: $\mathbf{G}_{j+1}=\widetilde{\mathbf{H}}_j^{+}$ (if ZF) or
$\mathbf{G}_{j+1}=\left(\mathbf{H}_d^j\left(\mathbf{H}_d^j\right)^{\dagger}+\sigma_n^2\mathbf{I}_{n_R}\right)^{-1}\mathbf{H}_d^j$ (if MMSE)
Decide the symbol entry for detection: $k_{j+1}=\arg\min_{i\notin\{k_1,k_2,\cdots,k_j\}}\left\|\left(\mathbf{G}_{j+1}\right)_i\right\|^2$
Update: $j=j+1$
}
```

In the above iterative procedure, for the ZF case, the vector $(\mathbf{G}_j)_i$ denotes the *ith* row of the matrix $\mathbf{G}_j$, computed from the pseudoinverse of $\widetilde{\mathbf{H}}_{j-1}$, where the columns $k_1,\cdots,k_{j-1}$ are set to zero. However, for the MMSE case, the vector $(\mathbf{G}_j)_i$ denotes the the *ith* column of the matrix $\mathbf{G}_j$, computed from the MMSE equalizer of $\mathbf{H}_d^{j-1}$, where the columns $k_1,\cdots,k_{j-1}$ are set to zero. This is because these columns only related to the entries of $s_{k_1},\cdots,s_{k_j}$ which have already been estimated and cancelled. Thus, the system can be regarded as a degenerated V-BLAST system of Figure 12 where the transmitters $k_1,\cdots,k_j$ are removed.

3.2.3 BER Performance of various V-BLAST detection algorithms

The BER performance of the V-BLAST with ML, ZF, and MMSE algorithms is presented in the following. Both the transmitter and the receiver are equipped with four antennas, and a flat fading channel is used for simulation. Fig. 13 compares the BER performance of the ML detector with those of the ZF-based or MMSE-based V-BLAST algorithms without ordering, in which the detecting order is in sequence from the last transmit antenna to the first transmit antenna. As compared with the ZF-based and MMSE-based V-BLAST algorithms, the ML detector has better BER performance. However, the computation complexity of the ML detector exponentially increases as the modulation order or the number of transmit antennas increase. Instead, the ZF-based and MMSE-based V-BLAST algorithms, which are the linear detection methods combined with the interference cancellation methods, require much lower complexity than the ML detector, but their BER performance is significantly inferior to that of the ML detector.

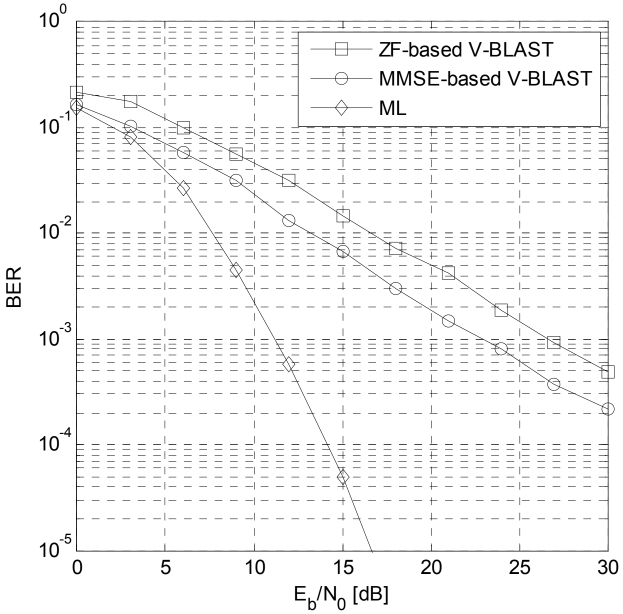


Fig. 13. BER performance of 4×4 V-BLAST systems without ordering

Fig. 14 and Fig. 15 demonstrate the BER performance of the ZF-based and the MMSE-based V-BLAST algorithm, respectively, with or without ordering. We can observe from these two figures that the V-BLAST algorithms with ordering can achieve better performance than that of the algorithms without ordering. An ordered successive interference suppression and cancellation method can effectively combat the error propagation problem to improve the BER performance with less complexity, although the ML detector still outperforms the ordering V-BLAST algorithms.

4. Beamforming

Beamforming is a promising signal processing technique used to control the directions for transmitting or receiving signals in spatial-angular domain (Godara, 1997). By adjusting beamforming weights, it can effectively concentrate its transmission or reception of desired signals at a particular spatial angle or suppress unwanted interference signals from other spatial angles. In this section, we will introduce the general concepts of beamforming techniques as well as some famous beamforming methods.

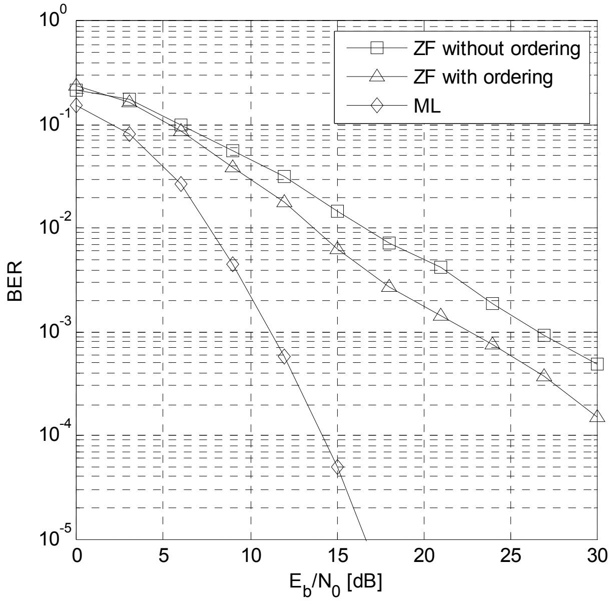


Fig. 14. BER performance of 4 × 4 V-BLAST systems using ZF-based V-BLAST algorithm with or without ordering

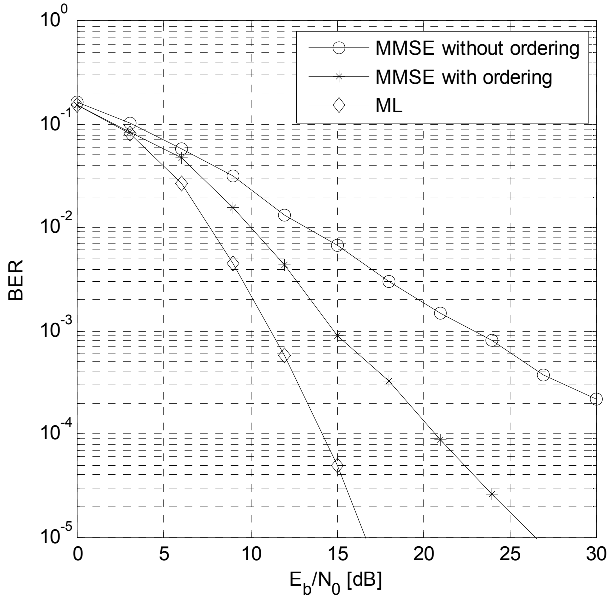


Fig. 15. BER performance of 4 × 4 V-BLAST systems using MMSE-based V-BLAST algorithm with or without ordering

4.1 Linear array and signal model

Fig. 16 depicts a linear array with L omni-directional and equi-spaced antennas, and the antenna spacing is set as d .

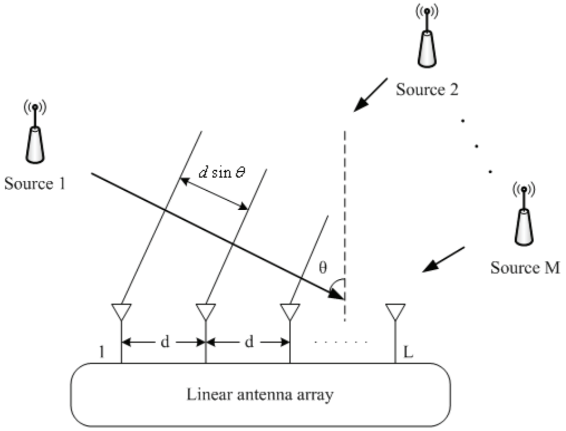


Fig. 16. Linear array

Let the first antenna element of the linear array be the reference point at the origin of the x-y coordinate system. Consider M uncorrelated far-field sources with the same central carrier frequency f_0 , for $i = 1,\dots,M$, and due to the far-field assumption, each source can be approximated as a plane wave when arriving at the linear array. Accordingly, the time arrival of the plane wave from the ith source in the direction of θ_i to the lth antenna is given by

$$\tau_l(\theta_i) = \frac{d}{c}(l-1)\sin\theta_i \tag{72}$$

where c is the speed of light. Denote $m_i(t)$ as the baseband signal of the ith source. By assuming that the bandwidth of the source is narrow enough, i.e., $m_i\left(t+\tau_l(\theta_i)\right) \cong m_i(t)$, the bandpass signal on the lth antenna contributed by the ith source is expressed as

$$\begin{aligned} b_i(t) &= \Re\mathrm{e}\left\{m_i\left(t+\tau_l(\theta_i)\right)e^{j2\pi f_0\left(t+\tau_l(\theta_i)\right)}\right\} \\ &\cong \Re\mathrm{e}\left\{m_i(t)e^{j2\pi f_0\left(t+\tau_l(\theta_i)\right)}\right\} \end{aligned} \tag{73}$$

Thus, the corresponding baseband equivalent signal on the lth antenna is given by

$$m_i(t)e^{j2\pi f_0\tau_l(\theta_i)} \tag{74}$$

Let $x_l(t)$ denote the baseband representation of the received signal on the lth antenna, including both M sources and noise on the lth antenna, and from (74), it is given by

$$x_l(t) = \sum_{i=1}^{M} m_i(t)e^{j2\pi f_0\tau_l(\theta_i)} + n_l(t). \tag{75}$$

where $n_l(t)$ is the spatially additive white Gaussian noise term on the lth antenna with zero mean and variance σ_n^2.

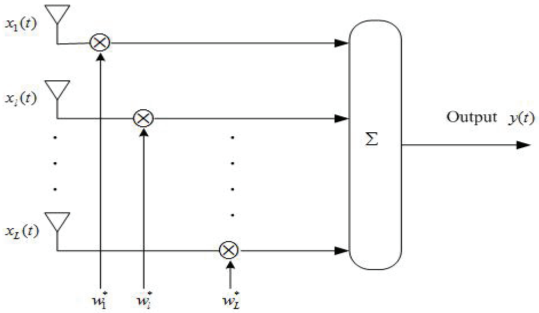


Fig. 17. Antenna beamforming

Fig. 17 shows the spatial signal processing of a beamforming array, in which complex beamforming weights w_l are applied to produce an output of linear combination of the received signals $x_l(t)$, expressed as

$$y(t)=\sum_{l=1}^{L} w_l^* x_l(t) \tag{76}$$

Define $\mathbf{w}=[w_1,w_2,\ldots,w_L]^T$ and $\mathbf{x}(t)=[x_1(t),x_2(t),\ldots,x_L(t)]^T$. From (75), we can express $\mathbf{x}(t)$ in a matrix-vector form as follows

$$\mathbf{x}(t)=\sum_{i=1}^{M} m_i(t)\mathbf{a}_i(\theta_i)+\mathbf{n}(t) \tag{77}$$

where $\mathbf{n}(t)=[n_1(t),n_2(t),\ldots,n_L(t)]^T$, and $\mathbf{a}_i(\theta_i)=\left[e^{j2\pi f_0\tau_1(\theta_i)},\ldots,e^{j2\pi f_0\tau_L(\theta_i)}\right]^T$ is known as the steering vector for the ith source. Then, we can rewrite (77) into a matrix notation, leading to a compact representation:

$$y(t)=\mathbf{w}^{\dagger}\mathbf{x}(t) \tag{78}$$

From (77) and (78), for a given beamforming vector, the mean output power is calculated as

$$P(\mathbf{w})=E\left[y(t)y^*(t)\right]=\mathbf{w}^{\dagger}\mathbf{R}\mathbf{w} \tag{79}$$

where $E[\bullet]$ denotes the expectation operator and $\mathbf{R}$ is the correlation matrix of the signal $\mathbf{x}(t)$, which is defined and given by

$$\mathbf{R}=E\left[\mathbf{x}(t)\mathbf{x}^{\dagger}(t)\right]=\sum_{i=1}^{M} p_i\mathbf{a}_i\mathbf{a}_i^{\dagger}+\sigma_n^2\mathbf{I} \tag{80}$$

where p_i is the power of the ith source, $\mathbf{I}$ is an identity matrix of size $L\times L$. In the following, we introduce three famous beamforming schemes to determine the complex beamforming weights.

4.2 Conventional beamforming

A conventional beamforming scheme is to form a directional beam by merely considering a single source. Since all its beamforming weights are set with an equal magnitude, this scheme is also named as delay-and-sum beamforming. Without loss of generality, assume that the targeted source for reception is in the direction of θ_0, and the beamforming weight vector is simply given by

$$\mathbf{w} = \frac{1}{L}\mathbf{a}(\theta_0) \tag{81}$$

Now we consider a communication environment consisting of only one signal source, and with this delay-and-sum beamforming scheme, the output signal after beamforming is given by

$$y(t) = \mathbf{w}^{\dagger}\mathbf{x}(t) = \mathbf{w}^{\dagger}\mathbf{a}(\theta_0)m_0(t) + \mathbf{w}^{\dagger}\mathbf{n}(t) \tag{82}$$

By assuming that the power of the source signal is equal to p_0, the post-output SNR after beamforming is then calculated as

$$SNR = \frac{\left\|\mathbf{w}^{\dagger}\mathbf{a}(\theta_0)m_0(t)\right\|^2}{E\left[\left\|\mathbf{w}^{\dagger}\mathbf{n}(t)\right\|^2\right]} = \frac{Lp_0}{\sigma_n^2} \tag{83}$$

In fact, the conventional beamforming can be regarded as an MRC-like scheme, as introduced in (Godara, 1997). It can be easily proved that for the case of a single source, the conventional beamforming scheme can provide the maximum output SNR. We can also observe that the output SNR in (83) is proportional to the number of antenna arrays, and as the number of antennas increases, it can facilitate to reduce the noise effect. However, when there are a number of signal sources which can interfere with each other, this conventional beamforming scheme does not have ability to suppressing interference effectively.

4.3 Null-steering beamforming

Consider a communication environment consisting of one desired signal source with an angle of arrival θ_0 and multiple interfering signal sources with angles of arrival θ_i, for $i = 1,\ldots,M$. To effectively mitigate the mutual interference, a null-steering beamforming scheme can be designed to null out unwanted signals from some interfering sources with known directions. The null-steering beamforming scheme is also named as ZF beamforming. As its name suggested, the design idea is to form beamforming weights with unity response in the desired source direction θ_0, while create multiple nulls in the interfering source directions θ_i, for $i = 1,\ldots,M$. Now assume that the steering vector for the desired and interfering signal sources are respectively denoted by $\mathbf{a}_0$ and $\mathbf{a}_i$, for $i = 1,\ldots,M$. Then, the beamforming weights can be designed by solving the following equations (D'Assumpcao & Mountford, 1984):

$$\begin{aligned} &\mathbf{w}^{\dagger}\mathbf{a}_0 = 1 \\ &\mathbf{w}^{\dagger}\mathbf{a}_i = 0, \text{ for } i = 1,\cdots,M \end{aligned} \tag{84}$$

Using the matrix notation, we can rewrite (84) as follows

$$\mathbf{w}^{\dagger}\mathbf{A} = \mathbf{e}_1^T \tag{85}$$

where $\mathbf{A}$ is a matrix containing $M+1$ steering vectors, i.e., $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_M]$, and we define $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$. If the total number of desired and interfering sources is equal to the number of antennas, $\mathbf{A}$ is an invertible square matrix as long as $\theta_i \neq \theta_j$, for all i, j. As a result, the solution for the beamforming weights is given by

$$\mathbf{w}^{\dagger} = \mathbf{e}_1^T \mathbf{A}^{-1} \tag{86}$$

Otherwise, the solution is obtained via taking the pseudo inverse of $\mathbf{A}$. The solution for the beamforming weights becomes

$$\mathbf{w}^{\dagger} = \mathbf{e}_1^T \mathbf{A}^{\dagger}\left(\mathbf{A}\mathbf{A}^{\dagger}\right)^{-1} \tag{87}$$

It is noted that although this beamforming scheme has nulls in the directions of interfering sources, it is not designed to maximum the output SNR. Hence, it is not an optimal one from the viewpoint of maximizing the output SNR.

4.4 Optimal beamforming

Although the null-steering beamforming scheme can deal with the interference problem by nulling out unwanted signals, it suffers from two critical problems. One is that the null-beamforming scheme requires the knowledge of interfering source directions, which is not easily acquired in practice. The other is that the output SNR is not the maximum, even though this scheme is quite simple. To overcome these two problems in the environment comprising of one desired signal source with the steering vector $\mathbf{a}_0$ and M interfering sources with the steering vectors $\mathbf{a}_i$, for $i = 1, \ldots, M$, from (79), we can derive the optimal beamforming weights by minimizing the total output power, under the constraint that the output gain in the direction of desired signal source is equal to one; that is

$$\begin{aligned} &\min_{\mathbf{w}} \quad \mathbf{w}^{\dagger}\mathbf{R}\mathbf{w} \\ &\textit{subject to} \;\; \mathbf{w}^{\dagger}\mathbf{a}_0 = 1 \end{aligned} \tag{88}$$

where $\mathbf{R} = \sum_{i=0}^{M} p_i \mathbf{a}_i \mathbf{a}_i^{\dagger} + \mathbf{R}_n$, p_i is the signal power of the corresponding source, and $\mathbf{R}_n$ is the covariance matrix of the spatial noise. Here, we consider a more general case where the noise is not necessary spatial white. This optimal beamforming scheme is also known as minimum variance distortionless response (MVDR) beamforming. The Lagrange multiplier can be applied to solve this optimization problem, and the Lagrange function associated with (88) is given by

$$L(\mathbf{w}, v) = \mathbf{w}^{\dagger}\mathbf{R}\mathbf{w} + v\left(\mathbf{w}^{\dagger}\mathbf{a}_0 - 1\right) = \mathbf{w}^{\dagger}\left(\mathbf{R}\mathbf{w} + v\mathbf{a}_0\right) - v \tag{89}$$

where v is the Lagrange multiplier. By applying the Karush-Kuhn-Tucker (K.K.T.) conditions, the sufficient conditions to reach the optimal solution are given by

$$\nabla_{\mathbf{w}} L(\mathbf{w}, v) = 2\mathbf{R}\mathbf{w} + v\mathbf{a}_0 = 0 \tag{90}$$

$$\mathbf{w}^{\dagger}\mathbf{a}_0 = 1 \tag{91}$$

We can further rewrite (90) as

$$\mathbf{w}^{\dagger} = -\frac{v^*}{2}\left(\mathbf{R}^{-1}\mathbf{a}_0\right)^{\dagger} \tag{92}$$

Since $\mathbf{R}$ is a Hermitian matrix, i.e., $\mathbf{R} = \mathbf{R}^{\dagger}$, by taking (91) into (92), we have

$$-\frac{v^*}{2} = \frac{1}{\mathbf{a}_0^{\dagger}\mathbf{R}^{-1}\mathbf{a}_0} \tag{93}$$

From (92) and (93), the optimal beamforming weights (Vural, 1975; Applebaum, 1976) are therefore given by

$$\mathbf{w} = \frac{\mathbf{R}^{-1}\mathbf{a}_0}{\mathbf{a}_0^{\dagger}\mathbf{R}^{-1}\mathbf{a}_0} \tag{94}$$

In general, the corelation matrix $\mathbf{R} = E\left[y(t)y^*(t)\right]$ can be approximated by computing the time average of the received signal power over a period of time duration *T*, i.e., $\mathbf{R} \cong (1/T)\sum_{t=0}^{T-1}|y(t)|^2$. It is worthwhile to mention here that for this scheme, the output SNR can be maximized without requiring the knowledge on the directions of the interference sources. Now consider a particular case where only one desired signal source exists and noise is with the covariance matrix $\mathbf{R}_n$, i.e., $\mathbf{R} = p_0\mathbf{a}_0\mathbf{a}_0^{\dagger} + \mathbf{R}_n$. Use the inversion lemma, we have

$$\mathbf{R}^{-1} = \mathbf{R}_n^{-1} - \frac{p_0\mathbf{R}_n^{-1}\mathbf{a}_0\mathbf{a}_0^{\dagger}\mathbf{R}_n^{-1}}{1 + p_0\mathbf{a}_0^{\dagger}\mathbf{R}_n^{-1}\mathbf{a}_0} \tag{95}$$

After some straightforward manipulation, the optimal beamforming scheme is then degenerated to

$$\mathbf{w} = \frac{\mathbf{R}^{-1}\mathbf{a}_0}{\mathbf{a}_0^{\dagger}\mathbf{R}^{-1}\mathbf{a}_0} = \frac{\mathbf{R}_n^{-1}\mathbf{a}_0}{\mathbf{a}_0^{\dagger}\mathbf{R}_n^{-1}\mathbf{a}_0} \tag{96}$$

In practice, it is hard to know the noise correlation matrix $\mathbf{R}_n$; however, it is evident from (96) that one can use $\mathbf{R}$, which could be approximately obtained through the time average of the received signal power, instead of $\mathbf{R}_n$ for calculating the optimal beamforming weights in this single-user case. For another special case where noise is spatially white, the optimal beamforming scheme is then further degenerated to the conventional beamforming scheme by substituting $\mathbf{R}_n = \sigma_n^2\mathbf{I}$ into (96):

$$\mathbf{w} = \frac{\mathbf{R}_n^{-1}\mathbf{a}_0}{\mathbf{a}_0^{\dagger}\mathbf{R}_n^{-1}\mathbf{a}_0} = \frac{1}{L}\mathbf{a}_0 \tag{97}$$

4.5 Performance of various beamforming techniques

We show some examples to evaluate the performance of these three beamforming schemes. The number of antenna and the carrier frequency of the sources are set as L = 20 and f_0 = 2.3 GHz, respectively. The antenna spacing is set to be half the wavelength. The number of sources is given by $M = 3$, and the directions of the desired and interfering signals are given by 20°, 5°, and 45°, respectively. Fig. 18 compares the angle responses of the conventional and optimal beamforming schemes. It is shown that the optimal beamforming scheme can efectively filter out the two interfering sources at the directions of 5° and 45°, as compared with the conventional beamforming scheme. Fig. 19 compares the angle response between the null-steering and the optimal beamforming schemes. Although the null-steering beamforming scheme can form deep nulls in the directions of the interfering sources, the performance of filtering out the interference sources degrades dramatically when there is small discrepancy between the true and esimated directions. In other words, the null-steering scheme requires extremely accurate estimation on the directions of interfering sources. Since the optimal beamforming does not require the knowledge of the directions of interfering sources; in fact, it can be performed by using the received signals to calcuate the correlation matrix in (90) instead, the optimal beamforming has more robust performance than the null-steering one.

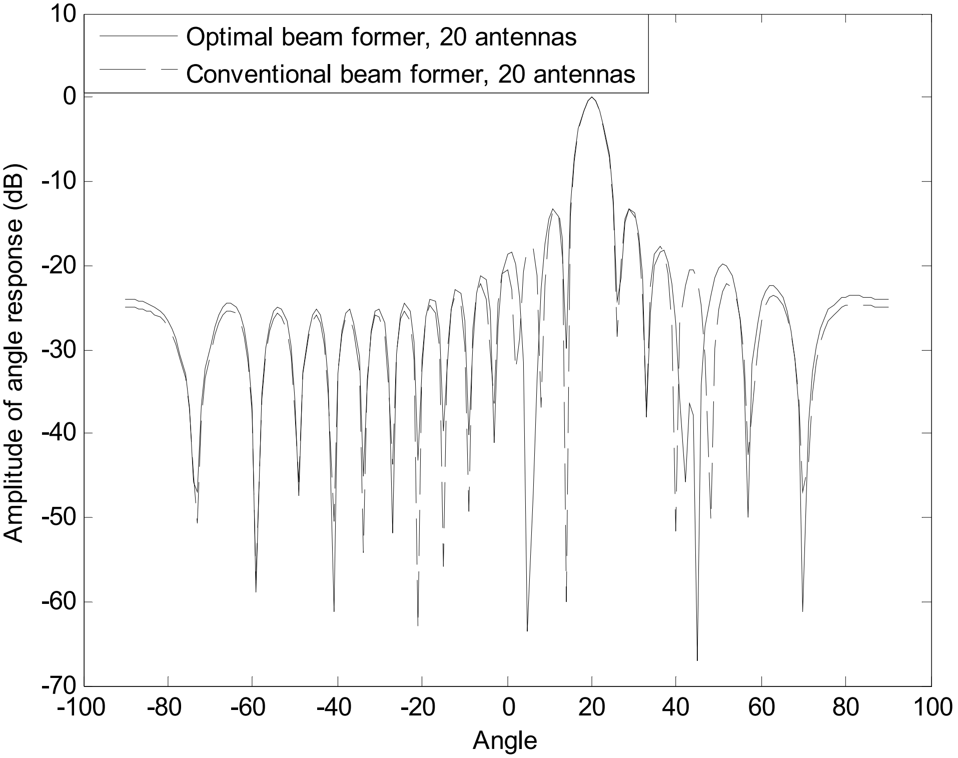


Fig. 18. Angle responses of optimal and conventional beamforming schemes

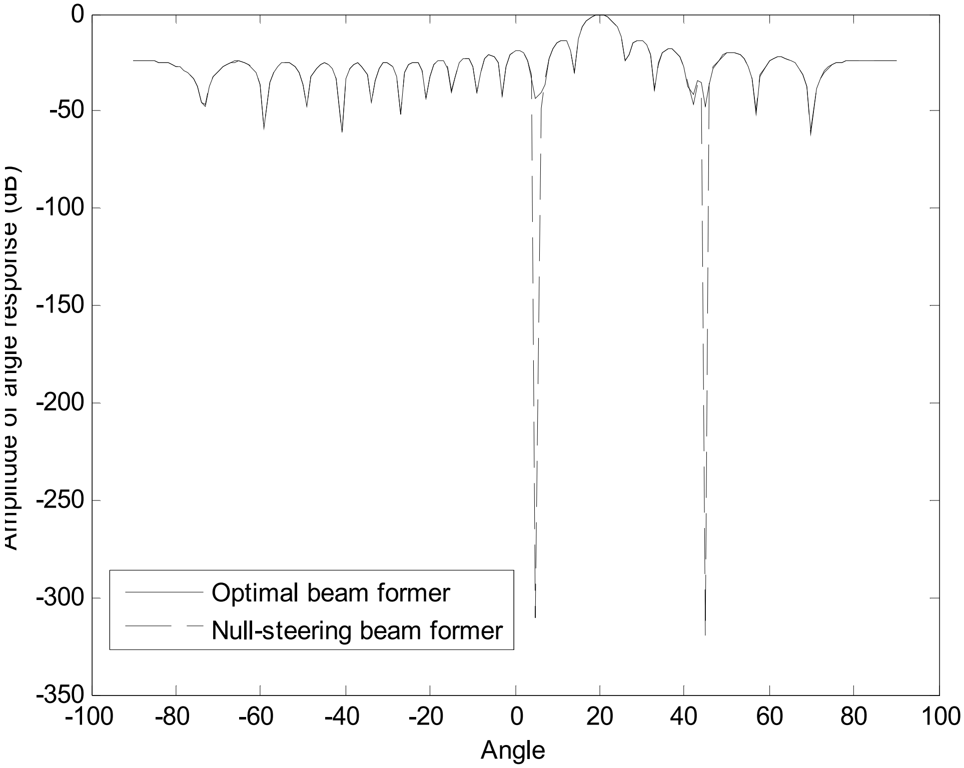


Fig. 19. Angle responses of optimal and null-steering schemes

5. References

Abadir, K. M. & Magnus, J. R. (2006). *Matrix Algebra*, Cambridge University Press, ISBN 978-0-521-53746-9, New York, USA.

Applebaum, S. P. (1976). Adaptive Arrays, *IEEE Transactions on Antennas and Propagation*, Vol.24, No.5, (Sep 1976), pp.585–598, ISSN: 0018-926X.

D'Assumpcao, H. A. & Mountford, G. E. (1984). An Overview of Signal Processing for Arrays of Receivers, *Journal of the Institution of Engineers (Australia)*, Vol.4, pp.6–19, ISSN 0020-3319.

Foschini, G. J. (1996). Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-Element Antennas. *Bell Labs Technical Journal*, Vol.1, No.2, (Summer 1999), pp.41-59, ISSN 1538-7305.

Foschini, G. J. & Gans, M. J. (1998). On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas. *Wireless Personal Communications*, Vol.6, No.3, (March 1998), pp.311-335, ISSN 0929-6212.

Foschini, G. J., Golden, G. D., Valenzuela, R. A. & Wolniansky, P. W. (1999). Simplified Processing for High Spectral Efficiency Wireless Communication Employing Multi-

Element Arrays. *IEEE Journal on Selected Areas in Communications*, Vol.17, No.11, (Nov 1999), pp.1841-1852, ISSN 0733-8716.

Giannakis, G. B., Liu, Z., Zhuo, S. & Ma, X. (2006). *Space-Time Coding for Broadband Wireless Communications*, John Wiley, ISBN 978-0-471-21479-3, New Jersey, USA.

Godara, L. C. (1997). Application of Antenna Arrays to Mobile Communications. II. Beam-Forming and Direction-of-Arrival Considerations. *Proceedings of IEEE*, Vol.85, No.8, (Aug 1997), pp.1195-1245, ISSN 0018-9219.

Ku, M.-L. & Huang, C.-C. (2006). A Complementary Code Pilot-Based Transmitter Diversity Technique for OFDM Systems. *IEEE Transactions on Wireless Communications*, Vol. 5, No. 2, (March 2006), pp.504-508, ISSN 1536-1276.

Ku, M.-L. & Huang, C.-C. (2008). A Refined Channel Estimation Method for STBC/OFDM Systems in High-Mobility Wireless Channels. *IEEE Transactions on Wireless Communications*, Vol. 7, No. 11, (Nov 2008), pp.4312-4320, ISSN 1536-1276.

Lin, J.-C. (2009). Channel Estimation Assisted by Postfixed Pseudo-Noise Sequences Padded with Zero Samples for Mobile Orthogonal-Frequency-Division-Multiplexing Communications. *IET Communications*, Vol.3, No.4, (April 2009), pp.561-570, ISSN 1751-8628.

Lin, J.-C. (2009). Least-Squares Channel Estimation Assisted by Self-Interference Cancellation for Mobile Pseudo-Random-Postfix Orthogonal-Frequency-Division Multiplexing Applications. *IET Communications*, Vol.3, No.12, (Dec 2009), pp.1907-1918, ISSN 1751-8628.

Lo, T. K. Y. (1999). Maximum Ratio Transmission. *IEEE Transactions on Communications*, Vol. 47, No. 10, (Oct 1999), pp.1458-1461, ISSN 0090-6778.

Rappaport, T. S. (2002). *Wireless Communications: Principles and Practice* (2nd Edition), Prentice Hall, ISBN 0-13-042232-0, New Jersey, USA.

Simon, M. K. & Alouini, M. S. (1999). A Unified Performance Analysis of Digital Communication with Dual Selective Combining Diversity over Correlated Rayleigh and Nakagami-m Fading Channels. *IEEE Transactions on Communications*, Vol.47, No.1, (Jan 1999), pp.33-43, ISSN 0090-6778.

Tarokh, V., Seshadri, N. & Calderbank, A. R. (1998). Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction. Information Theory. *IEEE Transactions on Information Theory*, Vol.44, No.2, (Mar 1998), pp.744-765, ISSN 0018-9448.

Tarokh, V., Jafarkhani, H. & Calderbank, A. R. (1999). Space-Time Block Codes from Orthogonal Designs. *IEEE Transactions on Information Theory*, Vol.45, No.5, (Jul 1999), pp.1456-1467, ISSN 0018-9448.

Vural, A. M. (1975). An Overview of Adaptive Array Processing for Sonar Application. *Proceedings of Electronics and Aerospace Systems Conference, 1975 (EASCON'75)*, pp.34A–34M.

Wittneben, A. (1993). A New Bandwidth Efficient Transmit Antenna Modulation Diversity Scheme for Linear Digital Modulation. *Proceedings of IEEE Communications Conference*, pp.1630-1634, ISSN: 0-7803-0950-2, Geneva, Switzerland, May 23-26, 1993.

Wolniansky, P. W., Foschini, G. J., Golden, G. D. & Valenzuela, R. A. (1998). V-BLAST: An Architecture for Realizing Very High Data Rates over The Rich-Scattering Wireless Channel. *Proceedings of URSI International Symposium on Signals, Systems, and Electronics*, pp.295-300, ISSN 0-7803-4900-8, Pisa, Italy, Sep 29-Oct 2, 1998.

5

Diversity Management in MIMO-OFDM Systems

Felip Riera-Palou and Guillem Femenias
Mobile Communications Group
University of the Balearic Islands
Spain

1. Introduction

Over the last decade, a large degree of consensus has been reached within the research community regarding the physical layer design that should underpin state-of-the-art and future wireless systems (e.g., IEEE 802.11a/g/n, IEEE 802.16e/m, 3GPP-LTE, LTE-Advanced). In particular, it has been found that the combination of multicarrier transmission and multiple-input multiple-output (MIMO) antenna technology leads to systems with high spectral efficiency while remaining very robust against the hostile wireless channel environment.

The vast majority of contemporary wireless systems combat the severe frequency selectivity of the radio channel using orthogonal frequency diversity multiplexing (OFDM) or some of its variants. The theoretical principles of OFDM can be traced back to (Weinstein & Ebert, 1971), however, implementation difficulties delayed the widespread use of this technique well until the late 80s (Cimini Jr., 1985). It is well-known that the combination of OFDM transmission with channel coding and interleaving results in significant improvements from an error rate point of view thanks to the exploitation of the channel frequency diversity (Haykin, 2001, Ch. 6). Further combination with spatial processing using one of the available MIMO techniques gives rise to a powerful architecture, MIMO-OFDM, able to exploit the various diversity degrees of freedom the wireless channel has to offer (Stuber et al., 2004).

1.1 Advanced multicarrier techniques

A significant improvement over conventional OFDM was the introduction of multicarrier code division multiplex (MC-CDM) by Kaiser (2002). In MC-CDM, rather than transmitting a single symbol on each subcarrier, as in conventional OFDM, symbols are code-division multiplexed by means of orthogonal spreading codes and simultaneously transmitted onto the available subcarriers. Since each symbol travels on more than one subcarrier, thus exploiting frequency diversity, MC-CDM offers improved resilience against subcarrier fading. This technique resembles very much the principle behind multicarrier code-division multiple access (MC-CDMA) where each user is assigned a specific spreading code to share a group of subcarriers with other users (Yee et al., 1993).

A more flexible approach to exploit the frequency diversity of the channel is achieved by means of group-orthogonal code-division multiplex (GO-CDM) (Riera-Palou et al., 2008). The idea behind GO-CDM, rooted in a multiple user access scheme proposed in (Cai et al., 2004), is to split suitably interleaved symbols from a given user into orthogonal groups, apply a spreading matrix on a per-subgroup basis and finally map each group to an orthogonal set of subcarriers. The subcarriers assigned to a group of symbols are typically chosen as separate as possible within the available bandwidth in order to maximise the frequency diversity gain. Note that a GO-CDM setup can be seen as many independent MC-CDM systems of lower dimension operating in parallel. This reduced dimension allows the use of optimum receivers for each group based on maximum likelihood (ML) detection at a reasonable computational cost. In (Riera-Palou et al., 2008), results are given for group dimensioning and spreading code selection. In particular, it is shown that the choice of the group size should take into account the operating channel environment because an exceedingly large group size surely leads to a waste of computational resources, and even to a performance degradation if the channel is not frequency-selective enough. Given the large variation of possible scenarios and equipment configurations in a modern wireless setup, a conservative approach of designing the system to perform satisfactorily in the most demanding type of scenario may lead to a significant waste of computational power, an specially scarce resource in battery operated devices. In fact, large constellation sizes (e.g., 16-QAM, 64-QAM) may difficult the application of GO techniques as the complexity of ML detection can become very high even when using efficient implementations such as the sphere decoder (Fincke & Pohst, 1985). In order to minimise the effects of a mismatch between the operating channel and the GO-CDM architecture, group size adaptation in the context of GO-CDM has been proposed in (Riera-Palou & Femenias, 2009), where it is shown that important complexity reductions can be achieved by dynamically adapting the group size in connection with the sensed frequency diversity of the environment.

1.2 Multiple antennae schemes

Multiple-antenna technology (i.e., MIMO) is the other main enabler towards high speed robust wireless networks. Whereas the use of multiple antennae at the receiver has been long applied as an effective measure to combat fading (see, e.g. (Simon & Alouini, 2005) and references therein), it is the application of multiple antenna at the transmitter side what revolutionised the wireless community. In particular, the linear increase in capacity achieved when jointly increasing the number of antennas at transmission and reception, theoretically forecasted in (Telatar, 1999), has spurred research efforts to effectively realize it in practical schemes. Among these practical schemes, three of them have achieved notable importance in the standardisation of modern wireless communications systems, namely, spatial division multiplexing (SDM), space-time block coding (STBC) and cyclic delay diversity (CDD). While in SDM (Foschini, 1996), independent data streams are sent from the different antennas in order to increase the transmission rate, in STBC (Alamouti, 1998; Tarokh et al., 1999) the multiple transmission elements are used to implement a space-time code targeting the improvement of the error rate performance with respect to that achieved with single-antenna transmission. In CDD (Wittneben, 1993) a single data stream is sent from all transmitter antennae with a different cyclic delay applied to each replica, effectively resulting as if the original stream was transmitted over a channel with increased frequency diversity.

1.3 Chapter objectives

The combination of GO-CDM and MIMO processing, termed MIMO-GO-CDM, results in a powerful and versatile physical layer able to exploit the channel variability in space and frequency. Nevertheless, the different MIMO processing schemes coupled with different degrees of frequency multiplexing (i.e., different group sizes) gives rise to a vast amount of combinations each offering a different operating point in the performance/complexity plane. Choosing an adequate number of Tx/Rx antennas, a specific MIMO scheme and the subcarrier grouping dimensions can be a daunting task further complicated when Tx and/or Rx antennas are correlated. To this end, it is desirable to have at hand closed-form analytical expressions predicting the performance of the different MIMO-GO-CDM configurations in order to avoid the need of (costly) numerical simulations.

This chapter has two main goals. The first goal is to present a unified BER analysis of the MIMO-GO-CDM architecture. In order to get an insight of the best possible performance this system can offer, attention is restricted to the case when ML detection is employed at the receiver. The analysis is general enough to incorporate the effects of channel frequency selectivity, Tx/Rx antenna correlations and the three most common different forms of spatial processing (SDM, STBC and CDD) in combination with GO-CDM frequential diversity. The analytical results are then used to explore the benefits of GO-CDM under different spatial configurations identifying the most attractive group dimensioning from a performance/complexity perspective. Based on the previous analysis, the second goal of this chapter is to devise effective reconfiguration strategies that can automatically and dynamically fix some of the parameters of the system, more in particular the group size of the GO-CDM component, in response to the instantaneous channel environment with the objective of optimising some pre-defined performance criteria (e.g., error rate, complexity, delay).

The rest of this chapter is organized as follows. Section 2 introduces the system model of a generic MIMO-GO-CDM system, paying special attention to the steps required to implement the frequency spreading and the MIMO processing. In Section 3 a unified BER analysis is presented for the case of ML detection. In light of this analysis, Section 4 explores reconfiguration strategies aiming at the optimisation of several critical parameters of the MIMO-GO-CDM architecture. Numerical results are presented in Section 6 to validate the introduced analytical and reconfiguration procedures. Finally, the main conclusions of this work are recapped in Section 7.

Notational remark: Vectors and matrices are denoted by bold lower and upper case letters, respectively. The superscripts *, ^{T} and ^{H} are used to denote conjugate, transpose and complex transpose (Hermitian), respectively, of the corresponding variable. The operation $vec(\boldsymbol{A})$ lines up the columns forming matrix $\boldsymbol{A}$ into a column vector. The symbols $\otimes$ and $\odot$ denote the Kronecker and element-by-element products of two matrices, respectively. Symbols $\boldsymbol{I}_k$ and $\mathbf{1}_{k\times l}$ denote the k-dimensional identity matrix and an all-ones $k \times l$ matrix, respectively. The symbol $\mathcal{D}(\boldsymbol{x})$ is used to represent a (block) diagonal matrix having $\boldsymbol{x}$ at its main (block) diagonal. The determinant of a square matrix $\boldsymbol{A}$ is represented by $|\boldsymbol{A}|$ whereas $\|\boldsymbol{x}\|^2 = \boldsymbol{x}\boldsymbol{x}^H$. Expression $\lceil a \rceil$ is used to denote the nearest upper integer of a. Finally, the Alamouti transform of a $K \times 2$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2]$ is defined as $\mathcal{A}(\boldsymbol{X}) \triangleq \left[-\boldsymbol{x}_2^*\ \boldsymbol{x}_1^*\right]$.

2. MIMO GO-CDM system model

We consider a MIMO multicarrier system with N_c data subcarriers, equipped with N_T and N_R transmit and receive antennas, respectively, and configured to transmit N_s $(\leq N_T)$ spatial data streams. Following the group-orthogonal design principles, the available subcarriers are split into $N_g = N_c/Q$ groups of Q subcarriers each. In the following subsections the transmitter, channel model and reception equation are described in detail.

2.1 Transmitter

As depicted in Fig. 1, incoming bits are split into N_s spatial streams, which are then processed separately. Bits on the zth stream are mapped onto a sequence $\boldsymbol{s}^z$ of symbols drawn from an

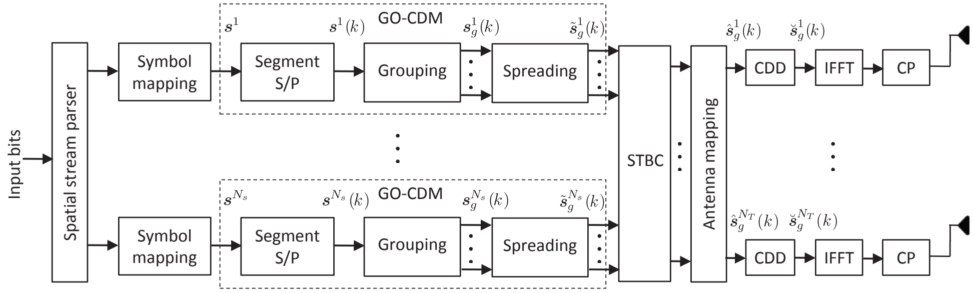


Fig. 1. Transmitter architecture for MIMO GO-CDM.

M-ary complex constellation (e.g. BPSK, M-QAM) with average normalized unit energy. The resulting N_s streams of modulated symbols $\{\boldsymbol{s}^z\}_{z=1}^{N_s}$ are then fed to the GO-CDM stage, which comprises three steps:

1. Segmentation of the incoming symbol stream in blocks of length N_c (i.e., eventual OFDM symbols), and serial to parallel conversion (S/P) resulting, over the kth OFDM symbol period, in $\boldsymbol{s}^z(k)$.
2. Arrangement of the symbols in the block into groups $\left\{\boldsymbol{s}^z_g(k)\right\}_{g=1}^{N_g}$, where $\boldsymbol{s}^z_g(k) = \left[s^z_{g,1}(k)\ldots s^z_{g,Q}(k)\right]^T$ represents an individual group.
3. Group spreading through a linear combination

$$\tilde{\boldsymbol{s}}^z_g(k) = \frac{1}{\sqrt{N_T}}\boldsymbol{C}\boldsymbol{s}^z_g(k), \tag{1}$$

   where $\boldsymbol{C}$ is a $Q \times Q$ orthonormal matrix, typically chosen to be a rotated Walsh-Hadamard matrix (Riera-Palou et al., 2008).

Before the usual OFDM modulation steps on each antenna (IFFT, guard interval appending and up-conversion), the grouped and spread symbols are processed in accordance with the MIMO transmission scheme in use as follows:

SDM ($N_s = N_T$**)** : In this case the blocks labeled in Fig. 1 as STBC and CDD are not used, and the spread symbols are directly supplied to the antenna mapping stage, which simply connects the incoming zth data stream to the ith transmit branch ($1 \leq i \leq N_T$), that is,

$$\check{\boldsymbol{s}}^i_g(k) = \hat{\boldsymbol{s}}^i_g(k) = \tilde{\boldsymbol{s}}^z_g(k). \tag{2}$$

STBC ($N_s = 1$**,** $N_T = 2$**)** : Two consecutive blocks of spread symbols, $\tilde{\boldsymbol{s}}^1_g(k)$ and $\tilde{\boldsymbol{s}}^1_g(k+1)$, are Alamouti-encoded on a per-subcarrier basis over two OFDM symbol periods,

$$\begin{aligned}\hat{\boldsymbol{s}}^1_g(k) &= \tilde{\boldsymbol{s}}^1_g(k), \qquad \hat{\boldsymbol{s}}^1_g(k+1) = -\left(\tilde{\boldsymbol{s}}^1_g(k+1)\right)^*,\\ \hat{\boldsymbol{s}}^2_g(k) &= \tilde{\boldsymbol{s}}^1_g(k+1),\ \hat{\boldsymbol{s}}^2_g(k+1) = \left(\tilde{\boldsymbol{s}}^1_g(k)\right)^*.\end{aligned} \tag{3}$$

In the antenna mapping stage, STBC-encoded streams are connected to two transmit branches, one for each symbol of the STBC code, that is,

$$\breve{\boldsymbol{s}}^{i}_{g}(k) = \hat{\boldsymbol{s}}^{i}_{g}(k). \tag{4}$$

CDD ($N_s = 1$) : In a pure CDD scheme, the same data stream is sent through N_T antennas with each replica being subject to a different cyclic delay Δ_i, typically chosen as $\Delta_i = \Delta_{i-1} + N_c/N_T$ with $\Delta_1 = 0$ (Bauch & Malik, 2006), resulting in transmitted symbols

$$\breve{s}^{i}_{g,q}(k) = \tilde{s}^{1}_{g,q}(k) \exp\left(-j2\pi d_q \Delta_i / N_c\right), \tag{5}$$

where d_q denotes the subcarrier index.

Hybrid schemes The analytical framework developed in this chapter can also be applied to hybrid systems combining SDM, STBC and/or CDD. Nevertheless, for brevity of presentation, the analysis to be developed next focuses on scenarios where only one of the mechanisms is used.

2.2 Channel model

The channel linking an arbitrary pair of Tx and Rx antennas is assumed to be time-varying and frequency-selective with an scenario-dependent power delay profile

$$S(\tau) = \sum_{l=0}^{P-1} \phi_l \delta(\tau - \tau_l), \tag{6}$$

where P denotes the number of independent paths of the channel and ϕ_l and τ_l denote the power and delay of the l-th path. It is assumed that the power delay profile is the same for all pairs of Tx and Rx antennas and that it has been normalized to unity (i.e., $\sum_{l=0}^{P-1} \phi_l = 1$). A single realization of the channel impulse response between Tx antenna i and receive antenna j at time instant t will then have the form

$$h^{ij}(t;\tau) = \sum_{l=0}^{P-1} h^{ij}_l(t) \delta(\tau - \tau_l), \tag{7}$$

where it will hold that $E\left\{\mid h^{ij}_l(t) \mid^2\right\} = \phi_l$. The corresponding frequency response can be expressed as

$$\bar{h}^{ij}(t;f) = \sum_{l=0}^{P-1} h^{ij}_l(t) \exp(-j2\pi f \tau_l), \tag{8}$$

which when evaluated at the N_c OFDM subcarriers yields

$$\bar{\boldsymbol{h}}^{ij}(t) = \left[\bar{h}^{ij}(t;f_0) \ldots \bar{h}^{ij}(t;f_{N_c-1})\right]^T. \tag{9}$$

In order to simplify the notation, assuming that the channel is static over the duration of a block (i.e., an OFDM symbol), the frequency response between Tx-antenna i and Rx-antenna j over the N_c subcarriers during the kth OFDM symbol can be expressed as

$$\bar{\boldsymbol{h}}^{ij}(k) = \left[\bar{h}^{ij}_0(k) \ldots \bar{h}^{ij}_{N_c-1}(k)\right]^T. \tag{10}$$

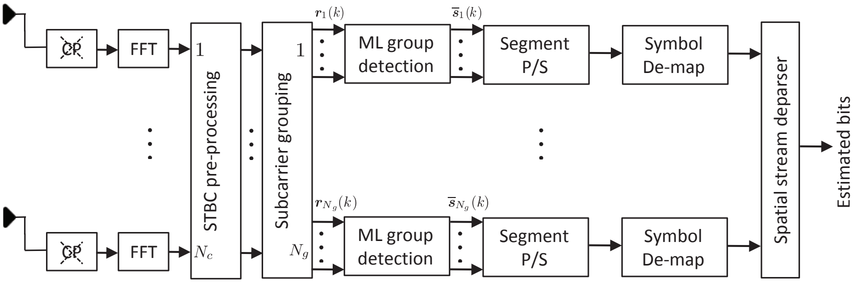


Fig. 2. Receiver architecture for MIMO-GO-CDM.

Since the subsequent analysis is mostly conducted on per-group basis, the channel frequency response for the gth group is denoted by

$$\bar{\boldsymbol{h}}_g^{ij}(k) = \left[\bar{h}_{g,1}^{ij}(k)\ldots\bar{h}_{g,Q}^{ij}(k)\right]^T, \tag{11}$$

with correlation matrix given by

$$\boldsymbol{\mathcal{R}}_{\boldsymbol{h}_g} = E\left\{\|\bar{\boldsymbol{h}}_g^{ij}(k)\|^2\right\} = E\left\{\bar{\boldsymbol{h}}_g^{ij}(k)\left(\bar{\boldsymbol{h}}_g^{ij}(k)\right)^H\right\}, \tag{12}$$

which is assumed to be constant over time, common for all pairs of Tx and Rx antennas and, provided that group subcarriers are chosen equispaced across the available bandwidth, common to all groups.

Now, considering the spatial correlation introduced by the transmit and receive antenna arrays, the spatially correlated channel frequency response for an arbitrary subcarrier q in group g can be expressed as (van Zelst & Hammerschmidt, 2002)

$$\boldsymbol{\mathcal{H}}_{g,q}(k) = \boldsymbol{\mathcal{R}}_{RX}^{1/2}\boldsymbol{H}_{g,q}(k)\left(\boldsymbol{\mathcal{R}}_{TX}^{1/2}\right)^T, \tag{13}$$

where $\boldsymbol{\mathcal{R}}_{RX}$ and $\boldsymbol{\mathcal{R}}_{TX}$ are, respectively, $N_R \times N_R$ and $N_T \times N_T$ matrices denoting the receive and transmit correlation, and

$$\boldsymbol{H}_{g,q}(k) = \begin{pmatrix} \bar{h}_{g,q}^{11}(k) & \ldots & \bar{h}_{g,q}^{1N_T}(k) \\ \vdots & & \vdots \\ \bar{h}_{g,q}^{N_R1}(k) & \ldots & \bar{h}_{g,q}^{N_RN_T}(k) \end{pmatrix}. \tag{14}$$

2.3 Receiver

As shown in Fig. 2, the reception process begins by removing the cyclic prefix and performing an FFT to recover the symbols in the frequency domain. After S/P conversion, and assuming ideal synchronization at the receiver side, the received samples for group g at the output of the FFT processing stage can be expressed in accordance with the MIMO transmission scheme in use as follows:

SDM and CDD: In these cases,

$$\boldsymbol{r}_g(k) = \text{vec}\left(\left[\boldsymbol{r}_{g,1}(k) \ldots \boldsymbol{r}_{g,Q}(k)\right]\right) = \boldsymbol{\mathcal{H}}_g(k)\breve{\boldsymbol{s}}_g(k) + \boldsymbol{v}_g(k), \tag{15}$$

where the $N_R Q \times N_T Q$ matrix

$$\boldsymbol{\mathcal{H}}_g(k) = \mathcal{D}\left(\left[\boldsymbol{\mathcal{H}}_{g,1}(k) \ldots \boldsymbol{\mathcal{H}}_{g,Q}(k)\right]\right), \tag{16}$$

represents the spatially and frequency correlated channel matrix affecting all symbols transmitted in group g, the $N_s Q$-long vector of transmitted (spread) symbols is formed as

$$\breve{\boldsymbol{s}}_g(k) = \text{vec}\left(\left[\breve{\boldsymbol{s}}_g^1(k) \ldots \breve{\boldsymbol{s}}_g^{N_T}(k)\right]^T\right), \tag{17}$$

and finally, $\boldsymbol{v}_g(k)$ is an $N_R Q \times 1$ vector representing the receiver noise, with each component being drawn from a circularly symmetric zero-mean white Gaussian distribution with variance σ_v^2.

STBC: As stated in (3), STBC encoding period $\eta = k/2$, with $k = 0, 2, 4, \ldots$, spawns two consecutive OFDM symbol periods, namely, the kth and $(k+1)$th symbol periods. Assuming that the channel coherence time is large enough to safely consider that $\boldsymbol{\mathcal{H}}_g(k+1) = \boldsymbol{\mathcal{H}}_g(k)$, then,

$$\begin{aligned} \tilde{\boldsymbol{r}}_g(k) &= \boldsymbol{\mathcal{H}}_g(k)\breve{\boldsymbol{s}}_g(k) + \boldsymbol{v}_g(k), \\ \tilde{\boldsymbol{r}}_g(k+1) &= \boldsymbol{\mathcal{H}}_g(k)\breve{\boldsymbol{s}}_g(k+1) + \boldsymbol{v}_g(k+1), \end{aligned} \tag{18}$$

and, therefore, we can define an equivalent received vector in STBC encoding period η as

$$\boldsymbol{r}_g(\eta) \triangleq \begin{bmatrix} \tilde{\boldsymbol{r}}_g(k) \\ \tilde{\boldsymbol{r}}_g^*(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mathcal{H}}_g(k) \\ \boldsymbol{\mathcal{H}}_g^{\mathcal{A}}(k) \end{bmatrix} \tilde{\boldsymbol{s}}_g(\eta) + \begin{bmatrix} \boldsymbol{v}_g(k) \\ \boldsymbol{v}_g^*(k+1) \end{bmatrix} \triangleq \tilde{\boldsymbol{\mathcal{H}}}_g(\eta)\tilde{\boldsymbol{s}}_g(\eta) + \tilde{\boldsymbol{v}}_g(\eta), \tag{19}$$

where

$$\boldsymbol{\mathcal{H}}_g^{\mathcal{A}}(k) \triangleq \mathcal{D}\left(\left[\mathcal{A}\left(\boldsymbol{\mathcal{H}}_{g,1}(k)\right) \ldots \mathcal{A}\left(\boldsymbol{\mathcal{H}}_{g,Q}(k)\right)\right]\right) \tag{20}$$

and

$$\tilde{\boldsymbol{s}}_g(\eta) \triangleq \text{vec}\left(\left[\tilde{\boldsymbol{s}}_g^1(k)\ \tilde{\boldsymbol{s}}_g^1(k+1)\right]^T\right). \tag{21}$$

In order to facilitate the unified performance analysis of the different MIMO strategies, it is more convenient to express the reception equation in terms of the original symbols rather than the spread ones. Thus, defining

$$\begin{aligned} \boldsymbol{s}_g(k) &= \tfrac{1}{\sqrt{N_T}}\text{vec}\left(\left[\boldsymbol{s}_g^1(k)\ \ldots\ \boldsymbol{s}_g^{N_s}(k)\right]^T\right) && \text{SDM} \\ \boldsymbol{s}_g(\eta) &= \tfrac{1}{\sqrt{2}}\text{vec}\left(\left[\boldsymbol{s}_g^1(k)\ \boldsymbol{s}_g^1(k+1)\right]^T\right) && \text{STBC} \\ \boldsymbol{s}_g(k) &= \tfrac{1}{\sqrt{N_T}}\boldsymbol{s}_g^1(k) && \text{CDD} \end{aligned} \tag{22}$$

it is straightforward to check that the symbols to be supplied to the IFFT processing step are given by,

$$\begin{aligned} \breve{\boldsymbol{s}}_g(k) &= \left(\boldsymbol{C} \otimes \boldsymbol{I}_{N_s}\right)\boldsymbol{s}_g(k) && \text{SDM} \\ \breve{\boldsymbol{s}}_g(k) &= \tilde{\boldsymbol{s}}_g(\eta) = \left(\boldsymbol{C} \otimes \boldsymbol{I}_2\right)\boldsymbol{s}_g(\eta) && \text{STBC} \\ \breve{\boldsymbol{s}}_g(k) &= \boldsymbol{E}_g^{\Delta}\left(\boldsymbol{C} \otimes \boldsymbol{1}_{N_T \times 1}\right)\boldsymbol{s}_g(k) && \text{CDD} \end{aligned}$$

with $\boldsymbol{E}_g^{\Delta} \triangleq \mathcal{D}\left(\left[\boldsymbol{E}_g^{\Delta 1} \ldots \boldsymbol{E}_g^{\Delta Q}\right]\right)$, where $\boldsymbol{E}_g^{\Delta q} = \mathcal{D}\left(\left[e^{-j2\pi d_q \Delta_1/N_c} \ldots e^{-j2\pi d_q \Delta_{N_T}/N_c}\right]\right)$ (Bauch & Malik, 2006). Furthermore, since processing takes place either on an OFDM symbol basis for SDM and CDD systems or on an STBC encoding period basis for STBC schemes, the indexes k and/or η can be dropped from this point onwards, allowing the reception equation to be expressed in general form as

$$\boldsymbol{r}_g = \boldsymbol{A}_g \boldsymbol{s}_g + \boldsymbol{\nu}_g$$

where

$$\boldsymbol{A}_g = \begin{cases} \boldsymbol{\mathcal{H}}_g \left(\boldsymbol{C} \otimes \boldsymbol{I}_{N_s}\right) & \text{SDM} \\ \tilde{\boldsymbol{\mathcal{H}}}_g \left(\boldsymbol{C} \otimes \boldsymbol{I}_2\right) & \text{STBC} \\ \boldsymbol{\mathcal{H}}_g \boldsymbol{E}_g^{\Delta} \left(\boldsymbol{C} \otimes \boldsymbol{1}_{N_T \times 1}\right) & \text{CDD} \end{cases}$$

and

$$\boldsymbol{\nu}_g = \begin{cases} \boldsymbol{v}_g & \text{for SDM/CDD} \\ \tilde{\boldsymbol{v}}_g & \text{for STBC} \end{cases} . \tag{23}$$

It should be noted that, regardless of the MIMO scheme and group dimension in use, the system matrix $\boldsymbol{A}_g$ has been normalised such that the SNR can be defined as $E_s/N_0 = 1/(2\sigma_v^2)$. Upon reception, all symbols in a group (for all streams in SDM and for both encoded OFDM symbols in STBC) are jointly estimated using an ML detection process. That is, the vector of estimated symbols in a group can be expressed as

$$\bar{\boldsymbol{s}}_g = \arg\min_{\boldsymbol{s}_g} \|\boldsymbol{A}_g \, \boldsymbol{s}_g - \boldsymbol{r}_g\|^2. \tag{24}$$

This procedure amounts to evaluate all the possible transmitted vectors and choosing the closest one (in a least-squares sense) to the received vector. Nevertheless, sphere detection (Fincke & Pohst, 1985) can be used for efficiently performing the exhaustive search required to implement the ML estimation.

3. Unified bit error rate analysis

3.1 BER analysis based on pairwise error probability

Using the well-known union bound (Simon et al., 1995), which is very tight for high signal-to-noise ratios, the bit error probability can be upper bounded as

$$P_b \leq \frac{1}{N_g N_Q M^{N_Q} \log_2 M} \sum_{g=1}^{N_g} \sum_{u=1}^{M^{N_Q}} \sum_{\substack{w=1 \\ w \neq u}}^{M^{N_Q}} P\left(\boldsymbol{s}_{g,u} \to \boldsymbol{s}_{g,w}\right) \mathcal{N}_b(\boldsymbol{s}_{g,u}, \boldsymbol{s}_{g,w}), \tag{25}$$

where,

$$N_Q = \begin{cases} Q\, N_s & \text{for SDM} \\ 2Q & \text{for STBC} \\ Q & \text{for CDD} \end{cases} . \tag{26}$$

The expression $P\left(\boldsymbol{s}_{g,u} \to \boldsymbol{s}_{g,w}\right)$, usually called the pairwise error probability (PEP), represents the probability of erroneously detecting the vector $\boldsymbol{s}_{g,w}$ when $\boldsymbol{s}_{g,u}$ was transmitted and

$\mathcal{N}_b(\boldsymbol{s}_{g,u}, \boldsymbol{s}_{g,w})$ is equal to the number of differing bits between vectors $\boldsymbol{s}_{g,u}$ and $\boldsymbol{s}_{g,w}$. To proceed further, the PEP conditioned on $\boldsymbol{A}_g$ can be shown to be (Craig, 1991)

$$\begin{aligned} P\left(\boldsymbol{s}_{g,u} \to \boldsymbol{s}_{g,w} | \boldsymbol{A}_g\right) &= \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{\|\boldsymbol{A}_g(\boldsymbol{s}_{g,u} - \boldsymbol{s}_{g,w})\|^2}{4\sigma_v^2}}\right) \\ &= \frac{1}{\pi}\int_0^{\pi/2} \exp\left(-\frac{\|\boldsymbol{A}_g(\boldsymbol{s}_{g,u} - \boldsymbol{s}_{g,w})\|^2}{4\sigma_v^2 \sin^2\phi}\right) d\phi. \end{aligned} \tag{27}$$

Now, defining the random variable $d^2_{g,uw} \triangleq \|\boldsymbol{A}_g(\boldsymbol{s}_{g,u} - \boldsymbol{s}_{g,w})\|^2$, the unconditional PEP can be expressed as

$$\begin{aligned} P\left(\boldsymbol{s}_{g,u} \to \boldsymbol{s}_{g,w}\right) &= \frac{1}{\pi}\int_0^{\pi/2}\int_{-\infty}^{+\infty} e^{-x/4\sigma_v^2 \sin^2\phi}\, p_{d^2_{g,uw}}(x)\, dx\, d\phi \\ &= \frac{1}{\pi}\int_0^{\pi/2} \mathcal{M}_{d^2_{g,uw}}\left(-\frac{1}{4\sigma_v^2 \sin^2\phi}\right) d\phi, \end{aligned} \tag{28}$$

where $p_x(\cdot)$ and $\mathcal{M}_x(\cdot)$ denote the probability density function (pdf) and moment generating function (MGF) of a random variable x, respectively.

Let us now define the error vector $\boldsymbol{e}_{g,uw} = \boldsymbol{s}_{g,u} - \boldsymbol{s}_{g,w}$. Using this definition, it can be shown that

$$d^2_{g,uw} \triangleq \|\boldsymbol{A}_g \boldsymbol{e}_{g,uw}\|^2 = \boldsymbol{\mathcal{H}}_g^H \boldsymbol{T}_{g,uw}^H \boldsymbol{T}_{g,uw} \boldsymbol{\mathcal{H}}_g, \tag{29}$$

where

$$\boldsymbol{\mathcal{H}}_g \triangleq \mathrm{vec}\left[\mathrm{vec}\left(\boldsymbol{\mathcal{H}}_{g,1}\right) \ldots \mathrm{vec}\left(\boldsymbol{\mathcal{H}}_{g,Q}\right)\right], \tag{30}$$

and $\boldsymbol{T}_{g,uw}$ can be expressed as

$$\boldsymbol{T}_{g,uw} = \begin{cases} \left[\left(\boldsymbol{1}_{Q\times 1} \otimes \boldsymbol{S}_{g,uw}\right) \odot \boldsymbol{\mathfrak{I}}_{Q,N_T}\right] \otimes \boldsymbol{I}_{N_R} & \text{SDM/CDD} \\ \left[\left(\boldsymbol{1}_{1\times Q} \otimes \boldsymbol{S}^T_{g,uw}\right) \odot \boldsymbol{\mathfrak{I}}^T_{Q,2}\right] \otimes \boldsymbol{I}_{2N_R} & \text{STBC} \end{cases} \tag{31}$$

with

$$\boldsymbol{S}_{g,uw} = \begin{cases} \boldsymbol{e}^T_{g,uw}\left(\boldsymbol{C}^T \otimes \boldsymbol{I}_{N_T}\right) & \text{SDM/STBC} \\ \boldsymbol{e}^T_{g,uw}\left(\boldsymbol{C}^T \otimes \boldsymbol{1}_{1\times N_T}\right) \boldsymbol{E}^T_{\Delta} & \text{CDD} \end{cases} \tag{32}$$

and $\boldsymbol{\mathfrak{I}}_{n,m} \triangleq \boldsymbol{I}_n \otimes \boldsymbol{1}_{1\times m}$. The expression of $d^2_{g,uw}$ reveals that it is a quadratic form in complex variables $\boldsymbol{\mathcal{H}}_g$, with MGF given by

$$\mathcal{M}_{d^2_{g,uw}}(s) = \left|\boldsymbol{I}_N - s\boldsymbol{G}_{g,uw}\right|^{-1}, \tag{33}$$

where N is equal to QN_R for the SDM and CDD schemes, and equal to $4QN_R$ for the STBC strategy. Furthermore,

$$\boldsymbol{G}_{g,uw} = \boldsymbol{T}_{g,uw} \boldsymbol{R}_g \boldsymbol{T}^H_{g,uw}, \tag{34}$$

with

$$\boldsymbol{R}_g = \boldsymbol{\mathcal{R}}_{\boldsymbol{h}_g} \otimes \boldsymbol{\mathcal{R}}_{TX} \otimes \boldsymbol{\mathcal{R}}_{RX}. \tag{35}$$

Now, let $\boldsymbol{\lambda}_{g,uw} = \{\lambda_{g,uw,1}, \ldots, \lambda_{g,uw,D_{g,uw}}\}$ denote the set of $D_{g,uw}$ distinct positive eigenvalues of $\boldsymbol{G}_{g,uw}$ with corresponding multiplicities $\boldsymbol{\alpha}_{g,uw} = \left\{\alpha_{g,uw,1}, \ldots, \alpha_{g,uw,D_{g,uw}}\right\}$. Using the results in (Femenias, 2004), the MGF of $d^2_{g,uw}$ can also be expressed as

$$\mathcal{M}_{d^2_{g,uw}}(s) = \prod_{d=1}^{D_{g,uw}} \frac{1}{(1 - s\lambda_{g,uw,d})^{\alpha_{g,uw,d}}} = \sum_{d=1}^{D_{g,uw}} \sum_{p=1}^{\alpha_{g,uw,d}} \frac{\kappa_{g,uw,d,p}}{(1 - s\lambda_{g,uw,d})^{p}} \tag{36}$$

where, using (Amari & Misra, 1997, Theorem 1), it can be shown that

$$\begin{aligned}\kappa_{g,uw,d,p} &= \frac{\lambda_{g,uw,d}^{p-\alpha_{g,uw,d}}}{(\alpha_{g,uw,d} - p)!} \frac{\partial^{\alpha_{g,uw,d}-p}}{\partial s^{\alpha_{g,uw,d}-p}} \left[\prod_{\substack{d'=1 \\ d' \neq d}}^{D_{g,uw}} \frac{1}{(1 - s\lambda_{g,uw,d'})^{\alpha_{g,uw,d'}}} \right] \Bigg|_{s=\frac{1}{\lambda_{g,uw,d}}} \\ &= \lambda_{g,uw,d}^{p-\alpha_{g,uw,d}} \sum_{\Phi} \prod_{\substack{d'=1 \\ d' \neq d}}^{D_{g,uw}} \frac{\lambda_{g,uw,d'}^{n_{d'}} \binom{\alpha_{g,uw,d'}+n_{d'}-1}{n_{d'}}}{\left(1 - \frac{\lambda_{g,uw,d'}}{\lambda_{g,uw,d}}\right)^{\alpha_{g,uw,d'}+n_{d'}}}\end{aligned} \tag{37}$$

with Φ being the set of nonnegative integers $\left\{n_1, \ldots, n_{d-1}, n_{d+1}, \ldots, n_{D_{g,uw}}\right\}$ such that $\sum_{d' \neq d} n_{d'} = \alpha_{g,uw,d} - p$, which allows (28) to be written as

$$\begin{aligned}P\left(\boldsymbol{s}_{g,u} \rightarrow \boldsymbol{s}_{g,w}\right) &= \frac{1}{\pi} \sum_{d=1}^{D_{g,uw}} \sum_{p=1}^{\alpha_{g,uw,d}} \kappa_{g,uw,d,p} \int_0^{\pi/2} \left(\frac{\sin^2\phi}{\sin^2\phi + \frac{\lambda_{g,uw,d}}{4\sigma_v^2}} \right)^p d\phi \\ &= \sum_{d=1}^{D_{g,uw}} \sum_{p=1}^{\alpha_{g,uw,d}} \kappa_{g,uw,d,p} \left(\frac{1 - \Omega\left(\frac{\lambda_{g,uw,d}}{4\sigma_v^2}\right)}{2} \right)^p \sum_{g=0}^{p-1} \binom{p-1+g}{g} \left(\frac{1 + \Omega\left(\frac{\lambda_{g,uw,d}}{4\sigma_v^2}\right)}{2} \right)^g,\end{aligned} \tag{38}$$

with $\Omega(c) = \sqrt{c/(1+c)}$. By substituting (38) into (25), a closed-form BER upper bound for an arbitrary power delay profile is obtained. It is later shown that this bound is tight, accurately matching the simulation results.

3.2 BER analysis based on PEP classes

Since there are many pairs $(\boldsymbol{s}_{g,u}, \boldsymbol{s}_{g,w})$ giving exactly the same PEP, it is possible to define a pairwise error class $\mathcal{C}(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c})$ as the set of all pairs $(\boldsymbol{s}_{g,u}, \boldsymbol{s}_{g,w})$ characterized by a common matrix $\boldsymbol{G}_{g,uw} = \boldsymbol{G}_{g,c}$ with $D_{g,c}$ distinct eigenvalues $\boldsymbol{\lambda}_{g,c} = \{\lambda_{g,c,1}, \ldots, \lambda_{g,c,D_{g,c}}\}$ with corresponding multiplicities $\boldsymbol{\alpha}_{g,c} = \{\alpha_{g,c,1}, \ldots, \alpha_{g,c,D_{g,c}}\}$ and therefore, a common PEP denoted by $\mathcal{P}(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c})$. A more insightful BER expression can then be obtained by using the PEP class notation, avoiding in this way the exhaustive computation of all the PEPs. Instead, the BER upper-bound can be found by computing the PEP for each class and weighing it using the number of elements in the class and the number of erroneous bits this class may induce. The BER upper bound can then be rewritten as

$$\begin{aligned}P_b &\leq \frac{1}{N_g N_Q M^{N_Q} \log_2 M} \\ &\quad \times \sum_{g=1}^{N_g} \sum_{\forall \mathcal{C}(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c})} \sum_{\mathcal{N}=1}^{N_Q \log_2 M} \mathcal{N}\, W(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c}, \mathcal{N}) \mathcal{P}(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c}),\end{aligned} \tag{39}$$

where $W(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c}, \mathcal{N})$ corresponds to the number of elements in the class $\mathcal{C}(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c})$ inducing $\mathcal{N}$ erroneous bits.

3.3 Asymptotic performance

Now, in order to gain further insight on the parameters affecting the BER performance, let us focus on the asymptotic case of large SNR. When $E_s/N_0 \to \infty$, the argument of the MGF in (28) also tends to infinity, and it can easily be shown that when $s \to \infty$ the MGF in (36) can be approximated by

$$\mathcal{M}_{d^2_{g,uw}}(s) \simeq \frac{1}{\left(\prod_{d=1}^{D_{g,uw}} \lambda_{g,uw,d}^{\alpha_{g,uw,d}}\right)(-s)^{\sum_{d=1}^{D_{d,uw}} \alpha_{g,uw,d}}}, \tag{40}$$

allowing the asymptotic PEP of the different classes to be expressed as

$$\mathcal{P}_{\text{asym}}\left(D_{g,c}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c}\right) = \frac{1}{\pi}\int_0^{\pi/2} \frac{(4\sigma_v^2 \sin^2\phi)^{\tilde{D}_{g,c}}}{\prod_{d=1}^{D_{g,c}} \lambda_{g,c,d}^{\alpha_{g,c,d}}} d\phi = \frac{(2\tilde{D}_{g,c})!}{2\tilde{D}_{g,c}!^2} \frac{(E_s/N_0)^{-\tilde{D}_{g,c}}}{\prod_{d=1}^{\tilde{D}_{\min}} \lambda_{g,c,d}^{\alpha_{g,c,d}}}, \tag{41}$$

where $\tilde{D}_{g,c} = \sum_{d=1}^{D_{g,c}} \alpha_{g,c,d}$ is the rank of the matrix-defining class $\boldsymbol{G}_{g,c}$. From (41) it is clear that the probability of error will be mainly determined by the groups and classes whose matrices

$$\boldsymbol{G}_{g,c} = \boldsymbol{G}_{g,c}^{\min} \triangleq \boldsymbol{T}_{g,c}^{\min} \boldsymbol{R}_{g}^{\min} \left(\boldsymbol{T}_{g,c}^{\min}\right)^H \tag{42}$$

have the smallest common rank, denoted by

$$\tilde{D}_{\min} = \text{rank}(\boldsymbol{G}_{g,c}^{\min}) = \text{rank}\left(\boldsymbol{T}_{g,c}^{\min} \boldsymbol{R}_{g}^{\min} \left(\boldsymbol{T}_{g,c}^{\min}\right)^H\right), \tag{43}$$

allowing the asymptotic BER to be written as

$$P_b \leq \sum_{g=1}^{N_g} \sum_{\forall \mathcal{C}(\tilde{D}_{\min}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c})} \sum_{\mathcal{N}=1}^{N_Q \log_2 M} \mathcal{N} \frac{(2\tilde{D}_{\min})!}{2(\tilde{D}_{\min}!)^2} \frac{W(\tilde{D}_{\min}, \boldsymbol{\lambda}_{g,c}, \boldsymbol{\alpha}_{g,c}, \mathcal{N})\,(E_s/N_0)^{-\tilde{D}_{\min}}}{N_g N_Q\, M^{N_Q} \log_2 M \prod_{d=1}^{\tilde{D}_{\min}} \lambda_{g,c,d}^{\alpha_{g,c,d}}}. \tag{44}$$

In light of (44), the asymptotic BER minimisation is achieved by maximising the minimum group/class rank $\tilde{D}_{\min}$ and the eigenvalue product of all the groups/classes with rank $\tilde{D}_{\min}$. In the following, only the maximization of $\tilde{D}_{\min}$ (i.e., maximisation of the diversity order) is pursued since the maximization of the product of eigenvalues is far more difficult as it involves the simultaneous optimization of all the eigenvalue products in the groups/classes with rank $\tilde{D}_{\min}$.

On the rank of $\boldsymbol{T}_{g,c}^{\min}$: As stated in (Cai et al., 2004; Riera-Palou et al., 2008), choosing the subcarriers for a group equispaced across the whole bandwidth minimizes subcarrier correlation allowing the optimization of the system performance if an adequate family of spreading codes is properly selected. To this end, rotated spreading transforms have been proposed for multicarrier systems in (Bury et al., 2003) where it is shown that the often used Walsh-Hadamard codes lead to poor diversity gains when employed to perform the frequency spreading. This can be explained by the fact that for certain symbol blocks the energy is concentrated on one single subcarrier and, thus,

$$\text{rank}\left(\boldsymbol{T}_{g,c}^{\min}\right) = \begin{cases} N_R & \text{SDM} \\ N_T\, N_R & \text{STBC/CDD.} \end{cases} \tag{45}$$

A deep fade on this subcarrier dramatically raises the probability of error in the detection process, regardless of the state of all other subcarriers, limiting in this way the achievable diversity order (asymptotic BER slope). A similar effect can be observed when using other spreading sequences such as those based on the discrete Fourier transform (DFT). As pointed out in (Bury et al., 2003), a spreading that has the potential to maximize the diversity order can be found by applying a rotation to the columns of the conventional spreading matrix $\boldsymbol{C}_{\text{conv}}$ as $\boldsymbol{C} = \boldsymbol{C}_{\text{conv}}\mathcal{D}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = [\theta_1 \ldots \theta_Q]$ with each θ_q denoting the chip-specific rotation, which in the proposed scheme is given by

$$\theta_q = \exp\left(\frac{j2\pi(q-1)}{Q\,\Theta}\right),$$

with Θ being constellation dependent and selected so as to make $2\pi/\Theta$ the minimum angle producing a rotation of the transmit symbol alphabet onto itself (e.g., $\Theta = 2$ for BPSK, $\Theta = 4$ for MQAM). This indicates that while using conventional Walsh-Hadamard spreading no frequency diversity gain will be achieved, the rotated spreading has the potential (depending on the channel correlation matrix $\boldsymbol{R}_g$) to attain a frequency diversity gain proportional to the number of subcarriers per group, common to all groups and classes. That is, when using optimally rotated spreading codes,

$$\text{rank}\left(\boldsymbol{T}_{g,c}^{\min}\right) = \begin{cases} Q\,N_R & \text{SDM} \\ Q\,N_T\,N_R & \text{STBC/CDD.} \end{cases} \tag{46}$$

On the rank of $\boldsymbol{R}_g^{\min}$: The correlation matrix $\boldsymbol{R}_g^{\min}$ can be expressed in general form as

$$\boldsymbol{R}_g^{\min} = \boldsymbol{\mathcal{R}}_{h_g}^{\min} \otimes \boldsymbol{\mathcal{R}}_{TX} \otimes \boldsymbol{\mathcal{R}}_{RX}, \tag{47}$$

and consequently (Petersen & Pedersen, 2008),

$$\text{rank}\left(\boldsymbol{R}_g^{\min}\right) = \text{rank}\left(\boldsymbol{\mathcal{R}}_{h_g}^{\min}\right)\text{rank}\left(\boldsymbol{\mathcal{R}}_{TX}\right)\text{rank}\left(\boldsymbol{\mathcal{R}}_{RX}\right). \tag{48}$$

Except for pathological setups exhibiting full spatial correlation between pairs of transmit or receive antennas (scenario not considered in this analysis), $\boldsymbol{\mathcal{R}}_{TX}$ and $\boldsymbol{\mathcal{R}}_{RX}$ are full rank matrices with $\text{rank}\left(\boldsymbol{\mathcal{R}}_{TX}\right) = N_T$ and $\text{rank}\left(\boldsymbol{\mathcal{R}}_{RX}\right) = N_R$, and therefore,

$$\text{rank}\left(\boldsymbol{R}_g^{\min}\right) = N_T\,N_R\text{rank}\left(\boldsymbol{\mathcal{R}}_{h_g}^{\min}\right). \tag{49}$$

Therefore, the maximum attainable frequency diversity order can be directly related to $\boldsymbol{\mathcal{R}}_{h_g}^{\min}$ and is given by the number of independent paths in the channel delay profile. If error performance is to be optimized, enough subcarriers per group need to be allocated to ensure that $\text{rank}\left(\boldsymbol{\mathcal{R}}_{h_g}^{\min}\right) = P$. In fact, defining the sampled channel order L as the channel delay spread in terms of chip (sampling) periods, it is shown in Cai et al. (2004) that the maximum rank of $\boldsymbol{\mathcal{R}}_{h_g}^{\min}$ is attained by setting the number of subcarriers per group to $Q = L + 1$. While this is a valuable design rule in channels with short delay spread, in most practical scenarios where L can be in the order of tens or even hundreds of samples, the theoretical number of subcarriers required to achieve full diversity would make the use of ML detection difficult even when using efficient search strategies (i.e., sphere decoding).

Moreover, very often maximum diversity would only be attained at unreasonably large E_s/N_0 levels.

In order to determine the number of subcarriers worth using in a given environment (i.e., a particular channel power delay profile), it is useful to use as reference the characteristics of the ideal case where all subcarriers in the group are fully uncorrelated (frequency domain *iid* channel). It is straightforward to see that, in this case, the frequency correlation matrix is given by $\boldsymbol{\mathcal{R}}_{h_g}^{\min} = \boldsymbol{I}_Q$, with rank $\left(\boldsymbol{\mathcal{R}}_{h_g}^{\min}\right) = Q$, and furthermore, it has only one non-zero eigenvalue $\lambda_{h_g,1} = 1$ with multiplicity $\alpha_{h_g,1} = Q$. Therefore, for any given MIMO configuration and a fixed number of subcarriers, the frequency domain *iid* channel results in the maximum frequency diversity order (Q) and will also lead to the minimum probability of error.

Since, for most realistic scenarios, setting the group size to guarantee full diversity ($Q = L + 1$) is unfeasible, we need to be able to measure what each additional subcarrier is contributing in terms of frequency diversity gain. Ideally, each additional subcarrier should bring along an extra diversity order, that is, an increase in rank $\left(\boldsymbol{\mathcal{R}}_{h_g}^{\min}\right)$ by one as it is indeed the case for uncorrelated channels. For correlated channels, however, this is often not the case and therefore to choose the group size it is useful to have some form of measure. A widely used tool in principal component analysis (Johnson & Wichern, 2002) to assess the *practical* dimensionality of a correlation matrix is the cumulative sum of eigenvalues (CSE) that, for the correlation matrix $\boldsymbol{\mathcal{R}}_{h_g}^{\min}$ with eigenvalues $\left\{\lambda_{h_g,q}\right\}_{q=1}^{Q}$, is defined as

$$\Psi(n) = \frac{\sum_{q=1}^{n} \lambda_{h_g,q}}{\sum_{q=1}^{Q} \lambda_{h_g,q}}. \tag{50}$$

For the frequency domain *iid* channel, $\Psi(n)$ is always a discrete linearly increasing function of n, and it can serve as a reference against which to measure the contribution of each subcarrier in arbitrary realistic channels.

As an example, suppose we are trying to determine the appropriate group size for models B and E from the propagation studies conducted in the definition of IEEE 802.11n (Erceg, 2003). Both models have been measured across a total bandwidth of 20 MHz with a channel sampling chip period of 10 ns. On one hand Model B is made of 11 paths and it has an *rms*-delay spread of 15 ns and very low frequency selectivity. On the other hand Model E corresponds to a channel with 38 paths (split in 4 clusters) with an *rms*-delay spread of 100 ns, resulting in large frequency selectivity. While Model B is representative of typical office indoor environments, Model E corresponds to large outdoor spaces such as airports or sport halls.

Figure 3 depicts the CSE for channel profiles B, E and the *iid* model, for different number of subcarriers ($Q = 2,\ 4,\ 8$ or 16) chosen equispaced across a bandwidth of 20 MHz. It can be inferred from the top left plot that when only two subcarriers are used per group ($Q = 2$), Models B and E behave qualitatively in a similar manner to the *iid* model and each of the subcarriers contributes in a significant way towards the achievement of the maximum diversity. When increasing the number of subcarriers (e.g., $Q = 4$, 8, 16), this no longer holds, notice how the CSE values for Model B quickly saturate and get farther apart from those of the *iid* channel, indicating that the additional subcarriers do not contribute substantially in increasing the frequency diversity order. For the case of Model E, a similar

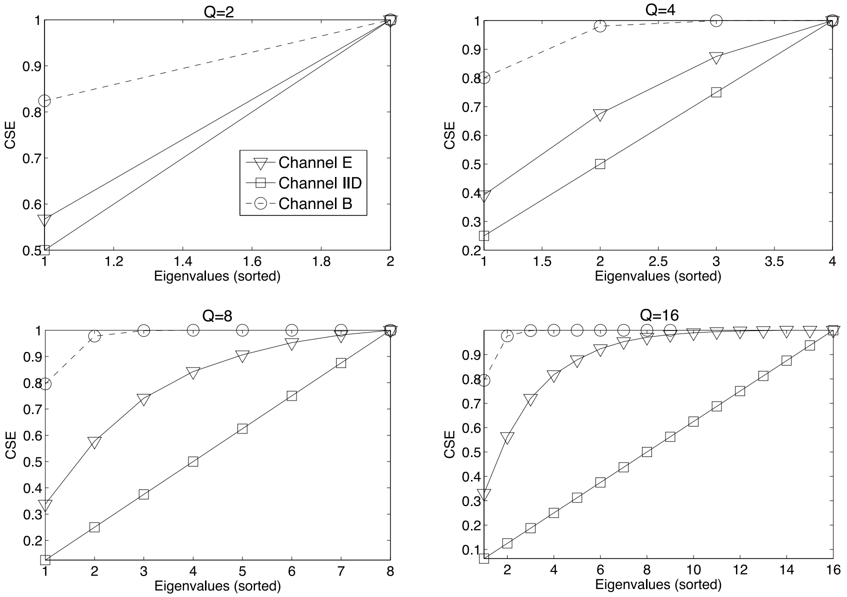


Fig. 3. Cumulative eigenvalue spread for Models B and E from (Erceg, 2003) and *iid* channel for different group sizes.

effect can be appreciated but to a much lesser extent, with the departure from the *iid* model being more evident for $Q = 16$ subcarriers. These results seem to indicate that, for Model B, $Q = 2$ would provide a good compromise between performance and detection complexity. In contrast, for channel E, $Q = 8$ would seem a more appropriate choice to fully exploit the channel characteristics. Notice that, according to the number of paths of each profile, Models B and E should attain diversity orders of 11 and 38, respectively. From the results in Fig. 3 it is obvious that far more moderate group sizes should be chosen in each case to operate in an optimal fashion from a diversity point of view at a reasonable (ML) detection complexity.

In conclusion, provided that scenarios with full spatial correlation are avoided, setting the number of subcarriers per group Q using the proposed CSE-based approach yields

$$\operatorname{rank}\left(\boldsymbol{R}_g^{\min}\right) = Q\, N_T\, N_R. \tag{51}$$

On the rank of $\boldsymbol{G}_{g,c}^{\min}$: Given an $m \times n$ matrix $\boldsymbol{A}$ and an $n \times p$ matrix $\boldsymbol{B}$, it holds that (Meyer, 2000)

$$\operatorname{rank}(\boldsymbol{A}) + \operatorname{rank}(\boldsymbol{B}) - n \leq \operatorname{rank}(\boldsymbol{AB}) \leq \min\left\{\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(\boldsymbol{B})\right\}. \tag{52}$$

Thus, using optimally rotated spreading codes and setting the number of subcarriers per group Q using the proposed CSE-based approach, provided that pathological scenarios with full spatial correlation are avoided, we can use (46) and (51) in (52) to show that the global diversity order for the analysed MIMO strategies is given by

$$\tilde{D}_{\min} = \operatorname{rank}\left(\boldsymbol{G}_{g,c}^{\min}\right) = \begin{cases} Q\, N_R & \text{SDM} \\ Q\, N_T\, N_R & \text{STBC/CDD.} \end{cases} \tag{53}$$

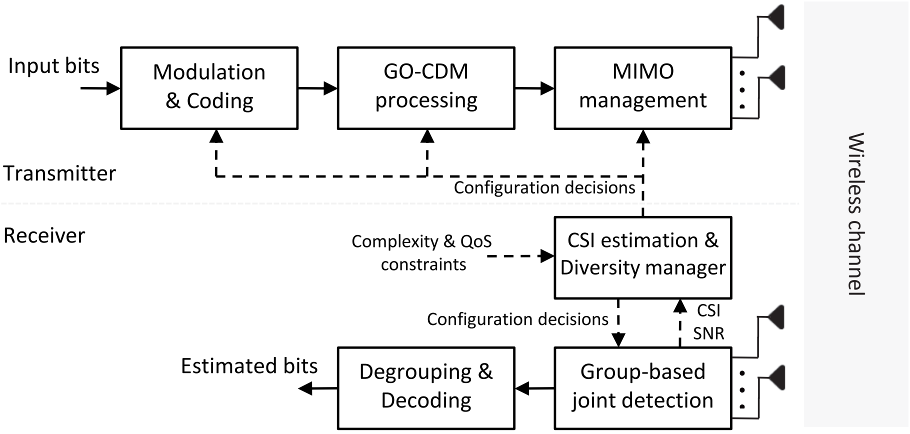


Fig. 4. Communication architecture for a MIMO-GO-CDM with group-size adaptation.

4. Reconfiguration strategies

It is clear from (44) and (53) that the (instantaneous) rank of the group frequency channel correlation matrix $\boldsymbol{\mathcal{R}}_{h_g}^{\min}$ determines the asymptotic diversity of a MIMO-GO-CDM system, and therefore, it can form the basis for a group size adaptation mechanism. Strictly speaking, the maximum possible rank of $\boldsymbol{\mathcal{R}}_{h_g}^{\min}$ is given by the number of independent paths in the channel profile. However, as shown in Subsection 3.3, very often the practical rank is far below this number as maximum diversity is only achieved at unrealistically low error rates. The adaptive group dimensioning scheme proposed next exploits this rank dependence to dynamically set the group size as a function of the channel response between all pairs of transmit and receive antennas. Figure 4 illustrates the architecture of the adaptive MIMO-GO-CDM system, where it can be appreciated that, in light of the acquired channel state information (CSI) and system constraints (complexity, QoS), the receiver determines the most appropriate group size to use and communicates this decision to the transmitter using a feedback channel. Note, as shown in Fig. 4, that CSI nd SNR information can also be used to determine the most appropriate modulation and coding scheme in conjunction with the GO-CDM dimensioning and MIMO mode selection. However this topic is beyond the scope of this chapter and in this work only fixed modulation and uncoded transmission modes are considered.

In order to perform the adaptive dimensioning of the GO-CDM component, the receiver requires an estimate $\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min}$ of the group frequency channel correlation matrix. An accurate estimate of the full correlation matrix $\boldsymbol{\mathcal{R}}_{h_g}^{\min}$ could be computed by means of time averaging over the frequency domain, however, in indoor/WLANs scenarios where channels tend to vary very slowly, this approach would require of many OFDM symbols to get an adequate estimate. Fortunately, only the group channel correlation matrix is required, thus simplifying the correlation estimation. Exploiting the grouping structure of GO-CDM-MIMO-OFDM and assuming the channel frequency response is a wide-sense stationary (WSS) process, it is possible to derive an accurate estimate $\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min}$ from the instantaneous CSI, provided the subcarriers in a given group have been chosen equispaced across the available bandwidth.

It is assumed that the group size to be determined is chosen from a finite set of possible values $\boldsymbol{Q} = \left\{Q^1, \ldots, Q^{\max}\right\}$ whose maximum, $Q^{\max}$, is limited by the maximum detection complexity the receiver can support. Suppose that at block symbol k the receiver acquires knowledge of the channel to form the frequency response $\bar{\boldsymbol{h}}^{ij}(k)$ over all N_c subcarriers. Now, using the maximum group size available, $Q^{\max}$, it is possible to form the frequency responses for all $N_g^{\min} = N_c/Q^{\max}$ groups, $\left\{\bar{\boldsymbol{h}}_1^{ij}(k), \ldots, \bar{\boldsymbol{h}}_{N_g^{\min}}^{ij}(k)\right\}$. Taking into account the WSS property it should hold that

$$E\left\{\bar{h}_{g,q}^{ij}(k)\bar{h}_{g,v}^{ij}(k)\right\} = E\left\{\bar{h}_{m,q}^{i'j'}(k)\bar{h}_{m,v}^{i'j'}(k)\right\}, \tag{54}$$

for all pairs of transmit and receive antennas (i,j) and (i',j') and any $q, v \in \{1, \ldots, Q^{\max}\}$, as the correlation among any two subcarriers should only depend on their separation, not their absolute position or the transmit/receive antenna pair. A group channel correlation matrix estimate from a single frequency response can now be formed averaging across transmit and receive antennas, and groups,

$$\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min} = \frac{1}{N_T N_R N_g^{\min}} \sum_{i=1}^{N_T} \sum_{j=1}^{N_R} \sum_{g=1}^{N_g^{\min}} \bar{\boldsymbol{h}}_g^{ij}(k)(\bar{\boldsymbol{h}}_g^{ij}(k))^H. \tag{55}$$

Using basic properties regarding the rank of a matrix, it is easy to prove that rank $\left(\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min}\right) \leq$ min $\left(N_g^{\min}, Q^{\max}\right)$, therefore, $N_g^{\min} = Q^{\max}$ maximises the range of possible group sizes using a single CSI shot. Let us denote the non-increasingly ordered positive eigenvalues of $\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min}$ by $\tilde{\boldsymbol{\Lambda}}_{h_g} = \left\{\tilde{\lambda}_{h_g,q}\right\}_{q=1}^{\tilde{Q}}$ where, owing to the deterministic character of $\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min}$, they can all be assumed to be different and with order one, and consequently, $\tilde{Q}$ represents the true rank of $\tilde{\boldsymbol{\mathcal{R}}}_{h_g}^{\min}$. For the purpose of adaptation, and based on the CSE criterion, a more flexible definition of rank is given as

$$\tilde{Q}_\epsilon = \min\left\{n \,:\, \Psi(n) = \frac{\sum_{q=1}^{n} \tilde{\lambda}_{h_g,q}}{\sum_{q=1}^{\tilde{Q}} \tilde{\lambda}_{h_g,q}} \geq 1 - \epsilon\right\}, \tag{56}$$

where $n \in \{1, \ldots, \tilde{Q}\}$ and ϵ is a small non-negative value used to set a threshold on the normalised CSE. Notice that $\tilde{Q}_\epsilon \to \tilde{Q}$ as $\epsilon \to 0$.

Since the group size Q represents the dimensions of an orthonormal spreading matrix $\boldsymbol{C}$, restrictions apply on the range of values it can take. For instance, in the case of (rotated) Walsh-Hadamard matrices, Q is constrained to be a power of two. The mapping of $\tilde{Q}_\epsilon$ to an allowed group dimension, jointly with the setting of ϵ, permits the implementation of different reconfiguration strategies, e.g.,

$$\text{Maximise performance} : Q = \arg\min_{\hat{Q} \in \boldsymbol{Q}} \{\hat{Q} \geq \tilde{Q}_\epsilon\} \tag{57a}$$

$$\text{Minimise complexity} : Q = \arg\min_{\hat{Q} \in \boldsymbol{Q}} \{|\hat{Q} - \tilde{Q}_\epsilon|\}. \tag{57b}$$

It is difficult to assess the feedback involved in this adaptive diversity mechanism as it depends on the dynamics of the underlying channel. The suggested strategy to implement

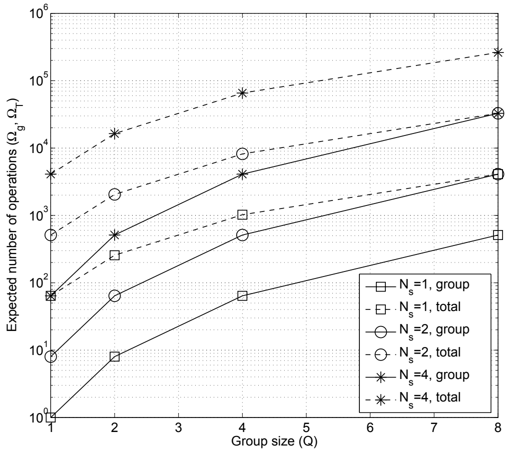


Fig. 5. Complexity as a function of group size (Q) for different number of transmitted streams.

this procedure is that the receiver regularly estimates the group channel rank and whenever a variation occurs, it determines and feeds back the new group dimension to the transmitter. In any case, the feedback information can be deemed insignificant as every update just requires of $\lceil \log_2 \mathcal{Q} \rceil$ feedback bits with $\mathcal{Q}$ denoting the cardinality of set $\boldsymbol{Q}$. Differential encoding of $\mathcal{Q}$ would bring this figure further down.

5. Computational complexity considerations

The main advantage of the group size adaptation technique introduced in the previous section is a reduction of computational complexity without any significant performance degradation. To gain some further insight, it is useful to consider the complexity of the detection process taking into account the group size in the GO-CDM component while assuming that an efficient ML implementation, such as the one introduced in (Fincke & Pohst, 1985), is in use. To this end, Vikalo & Hassibi (2005) demonstrated that the number of expected (complex) operations in an efficient ML detector operating at reasonable SNR levels is roughly cubic with the number of symbols jointly detected. That is, to detect one single group in a MIMO-GO-CDM system, $\Omega_g = \mathcal{O}(N_Q^3)$ operations are required.

Obviously, to detect all groups in the system, the expected number of required operations is given by $\Omega_T = \frac{N_c}{Q}\Omega_g$. Figure 5 depicts the expected per-group and total complexity for a system using $N_c = 64$ subcarriers, a set of possible group sizes given by $\{1, 2, 4, 8\}$ and different number of transmitted streams. Note that, in the context of this chapter, $N_s > 1$ necessarily implies the use of SDM. Importantly, increasing the group size from $Q = 1$ to $Q = 8$ implies an increase in the number of expected operations of more than two orders of magnitude, thus reinforcing the importance of rightly selecting the group size to avoid a huge waste in computational/power resources. Finally, it should be mentioned that for the STBC setup, efficient detection strategies exist that decouple the Alamouti decoding and GO-CDM

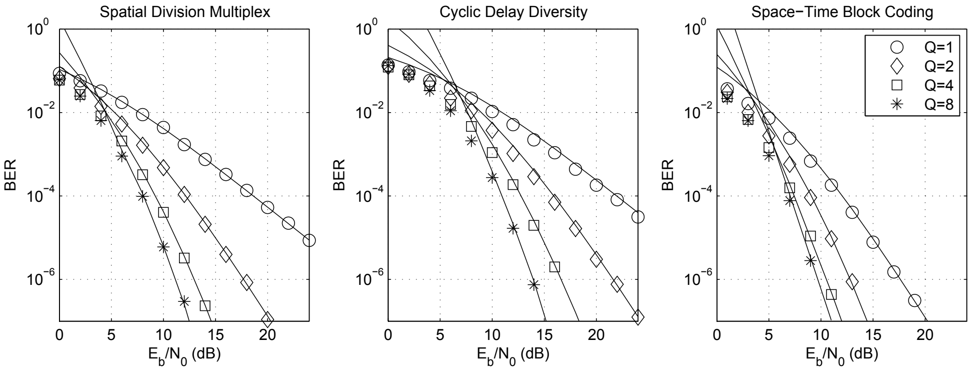


Fig. 6. Analytical (lines) and simulated (markers) BER for GO-CDM configured to operate in SDM (left), CDD (centre) and STBC (right) for different group sizes in Channel Profile E.

detection resulting in a simplified receiver architecture that is still optimum (Riera-Palou & Femenias, 2008).

6. Numerical results

In this section, numerical results are presented with the objective of validating the analytical derivations introduced in previous sections and also to highlight the benefits of the adaptive MIMO-GO-CDM architecture. The system considered employs $N_c = 64$ subcarriers within a $B = 20$ MHz bandwidth. These parameters are representative of modern WLAN systems such as IEEE 802.11n (IEEE, 2009). The GO-CDM technique has been applied by spreading the symbols forming a group with a rotated Walsh-Hadamard matrix of appropriate size. The set of considered group sizes is given by $\boldsymbol{Q} = \{1, 2, 4, 8\}$. This set covers the whole range of practical diversity orders for WLAN scenarios while remaining computationally feasible at reception. Note that a system with $Q = 1$ effectively disables the GO-CDM component. For most of the results shown next, Channel Profile E from (Erceg, 2003) has been used. Perfect channel knowledge is assumed at the receiver. Regarding the MIMO aspects, the system is configured with two transmit and two receive antennas ($N_T = N_R = 2$). As in (van Zelst & Hammerschmidt, 2002), the correlation coefficient between Tx (Rx) antennas is defined by a single coefficient ρ_{Tx} (ρ_{Rx}). Note that in order to make a fair comparison among the different spatial configurations, different modulation alphabets are used. For SDM, two streams are transmitted using BPSK whereas for STBC and CDD, a single stream is sent using QPSK modulation, ensuring that the three configurations achieve the same spectral efficiency.

Figure 6 presents results for SDM, CDD and STBC when transmit and receive correlation are set to $\rho_{Tx} = 0.25$ and $\rho_{Rx} = 0.75$, respectively. The first point to highlight from the three subfigures is the excellent agreement between simulated and analytical results for the usually relevant range of BERs ($10^{-3} - 10^{-7}$). It can also be observed the various degrees of influence exerted by the GO-CDM component depending on the particular spatial processing mechanism in use. For example, at a $P_b = 10^{-4}$, it can be observed that in SDM and CDD, the maximum group size considered ($Q = 8$) brings along SNR reductions greater than 10 dB when compared to the setup without GO-CDM ($Q = 1$). In contrast, in combination with STBC, the maximum gain offered by GO-CDM is just above 5 dB. The overall superior performance of STBC can be explained by the fact that it exploits transmit and receive

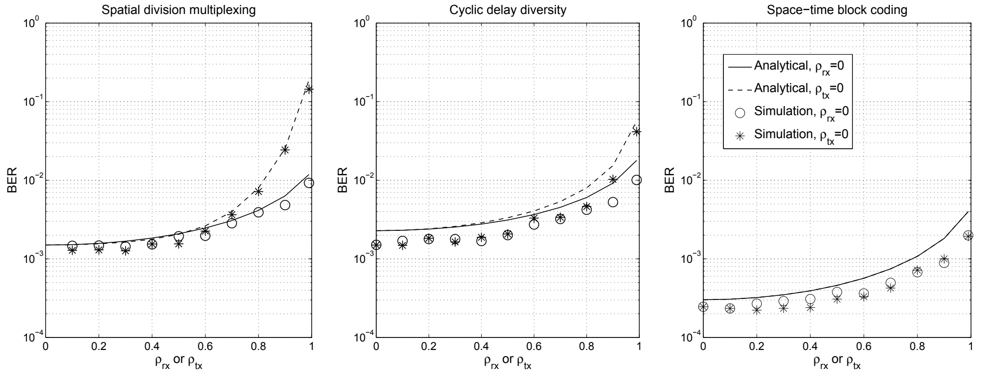


Fig. 7. Analytical (lines) and simulated (markers) BER for GO-CDM configured to operate in SDM (left), CDD (centre) and STBC (right) for different transmit/antenna correlation values.

diversity whereas in SDM there is no transmit diversity and in CDD, this is only exploited when combined with GO-CDM and/or channel coding.

Next, the effects of antenna correlation at either side of the communication link have been assessed for each of the MIMO processing schemes. To this end, the MIMO-GO-CDM system has been configured with $Q = 2$ and the SNR fixed to $E_s/N_0 = 10\ dB$. The antenna correlation at one side was set to 0 when varying the antenna correlation at the other end between 0 and 0.99. As seen in Fig. 7, a good agreement between analytical and numerical results can be appreciated. The small discrepancy between theory and simulation is mainly due to the use of the union bound, which always overestimates the true error rate. In any case, the theoretical expressions are able to predict the performance degradation due to an increased antenna correlation. Note that, in CDD and SDM, for low to moderate values ($0.0 - 0.7$), correlation at either end results in a similar BER degradation, however, for large values (> 0.7), correlation at the transmitter is significantly more deleterious than at the receiver. For the STBC scenario, analysis and simulation demonstrate that it does not matter which communication end suffers from antenna correlation as it leads to exactly the same results. This is because all symbols are transmitted and received through all antennas (Tx and Rx) and therefore equally affected by the correlation at both ends.

Finally, the performance of the proposed group adaptive mechanism has been assessed by simulation. The SNR has been fixed to $E_s/N_0 = 12\ dB$ and a time varying channel profile has been generated. This profile is composed of *epochs* of 10,000 OFDM symbols each. Within an epoch, an independent channel realisation for each OFDM symbol is drawn (quasi-static block fading) from the same channel profile. For visualisation clarity, the generating channel profile is kept constant for three consecutive epochs and then it changes to a different one. All channel profiles (A-F) from IEEE 802.11n (Erceg, 2003) have been considered. Results shown correspond to an SDM configuration.

The left plot in Fig. 8 shows the BER evolution for fixed and adaptive group size systems as the environment switches among the different channel profiles. The upper-case letter on the top of each plot identifies the particular channel profile for a given *epoch*. Each marker represents the averaged BER of 10,000 OFDM symbols. Focusing on the fixed group configurations it is easy to observe that a large group size does not always bring along a reduction in BER. For example, for Profile A (frequency-flat channel) there is no benefit in pursuing extra frequency

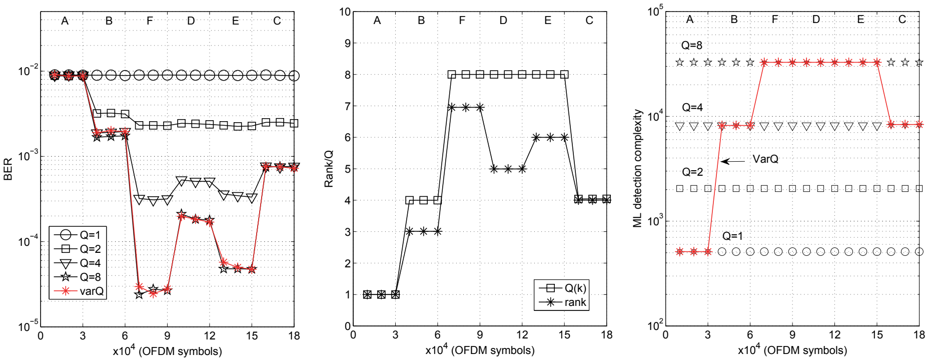


Fig. 8. Behaviour of fixed and adaptive MIMO GO-CDM-OFDM over varying channel profile using QPSK modulation at E_s/N_0=12 dB. $N_T = N_R = N_s = 2$ (SDM mode). Left: epoch-averaged BER performance. Middle: epoch-averaged rank/group size. Right: epoch-averaged detection complexity.

diversity at all. Similarly, for Profiles B and C there is no advantage in setting the group size to values larger than 4. This is in fact the motivation of the proposed MIMO adaptive group size algorithm denoted in the figure by *varQ*. It is clear from the middle plot in Fig. 8 that the proposed algorithm is able to adjust the group size taking into account the operating environment so that when the channel is not very frequency selective low Q values are used and, in contrast, when large frequency selectivity is sensed the group size dimension grows. Complementing the BER behaviour, it is important to consider the computational cost of the configurations under study. To this end the right plot in Fig. 8 shows the expected number of complex operations (see Section 5). In this plot it can be noticed the huge computational waste incurred, since there is no BER reduction, in the fixed group size systems with large Q when operating in channels with a modest amount of frequency-selectivity (A, B and C).

7. Conclusions

This chapter has introduced the combination of GO-CDM and multiple transmit antenna technology as a means to simultaneously exploit frequency, time and space diversity. In particular, the three most common MIMO mechanisms, namely, SDM, STBC and CDD, have been considered. An analytical framework to derive the BER performance of MIMO-GO-CDM has been presented that is general enough to incorporate transmit and receive antenna correlations as well as arbitrary channel power delay profiles. Asymptotic results have highlighted which are the important parameters that influence the practical diversity order the system can achieve when exploiting the three diversity dimensions. In particular, the channel correlation matrix and its effective rank, defined as the number of significant positive eigenvalues, have been shown to be the key elements on which to rely when dimensioning MIMO-GO-CDM systems. Based on this effective rank, a dynamic group size strategy has been introduced able to adjust the frequency diversity component (GO-CDM) in light of the sensed environment. This adaptive MIMO-GO-CDM has been shown to lead to important power/complexity reductions without compromising performance and it has the potential to incorporate other QoS requirements (delay, BER objective) that may result in further energy savings. Simulation results using IEEE 802.11n parameters have served to verify three facts. Firstly, MIMO-GO-CDM is a versatile architecture to exploit the different degrees of freedom the environment has to offer. Secondly, the presented analytical framework is able to accurately model the BER behaviour of the various MIMO-GO-CDM configurations. Lastly, the adaptive group size strategy is able to recognize the operating environment and adapt the system appropriately.

8. Acknowledgments

This work has been supported in part by MEC and FEDER under projects MARIMBA (TEC2005-00997/TCM) and COSMOS (TEC2008-02422), and a Ramón y Cajal fellowship (co-financed by the European Social Fund), and by Govern de les Illes Balears through project XISPES (PROGECIB-23A).

9. References

Alamouti, A. (1998). A simple transmit diversity technique for wireless communications, *IEEE JSAC* 16: 1451–1458.

Amari, S. & Misra, R. (1997). Closed-form expressions for distribution of sum of exponential random variables, *IEEE Trans. Reliability* 46(4): 519–522.

Bauch, G. & Malik, J. (2006). Cyclic delay diversity with bit-interleaved coded modulation in orthogonal frequency division multiple access, *IEEE Trans. Wireless Commun.* 8: 2092–2100.

Bury, A., Egle, J. & Lindner, J. (2003). Diversity comparison of spreading transforms for multicarrier spread spectrum transmission, *IEEE Trans. Commun.* 51(5): 774–781.

Cai, X., Zhou, S. & Giannakis, G. (2004). Group-orthogonal multicarrier CDMA, *IEEE Trans. Commun.* 52(1): 90–99.

Cimini Jr., L. (1985). Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing, *IEEE Transactions on Communications* 33(7): 665–675.

Craig, J. W. (1991). A new, simple and exact result for calculating the probability of error for two-dimensional signal constellations, *IEEE MILCOM'91 Conf. Rec.*, Boston, MA, pp. 25.5.1–25.5.5.

Erceg, V. (2003). Indoor MIMO WLAN Channel Models. doc.: IEEE 802.11-03/871r0, Draft proposal.

Femenias, G. (2004). BER performance of linear STBC from orthogonal designs over MIMO correlated Nakagami-m fading channels, *IEEE Trans. Veh. Technol.* 53(2): 307–317.

Fincke, U. & Pohst, M. (1985). Improved methods for calculating vectors of short length in a lattice, including a complexity analysis, *Math. Comput.* 44: 463–471.

Foschini, G. (1996). Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell Labs Technical Journal* 1(2): 41–59.

Haykin, S. (2001). *Communication Systems*, 4th edn, Wiley.

IEEE (2009). Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 5: Enhancements for Higher Throughput, *IEEE Std 802.11n-2009* .

Johnson, R. & Wichern, D. (2002). *Applied Multivariate Statistical Analysis*, fifth edn, Prentice Hall.

Kaiser, S. (2002). OFDM code-division multiplexing in fading channels, *IEEE Trans. Commun.* 50: 1266–1273.

Meyer, C. (2000). *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics (SIAM).

Petersen, K. B. & Pedersen, M. S. (2008). The matrix cookbook. Version 20081110.
URL: *http://www2.imm.dtu.dk/pubdb/p.php?3274*

Riera-Palou, F. & Femenias, G. (2008). Improving STBC performance in IEEE 802.11n using group-orthogonal frequency diversity, *Proc. IEEE Wireless Communications and Networking Conference*, Las Vegas (US), pp. 1–6.

Riera-Palou, F. & Femenias, G. (2009). OFDM with adaptive frequency diversity, *IEEE Signal Processing Letters* 16(10): 837 – 840.

Riera-Palou, F., Femenias, G. & Ramis, J. (2008). On the design of uplink and downlink group-orthogonal multicarrier wireless systems, *IEEE Trans. Commun.* 56(10): 1656–1665.

Simon, M. & Alouini, M. (2005). *Digital communication over fading channels*, Wiley-IEEE Press.

Simon, M., Hinedi, S. & Lindsey, W. (1995). *Digital communication techniques: signal design and detection*, Prentice Hall PTR.

Stuber, G., Barry, J., Mclaughlin, S., Li, Y., Ingram, M. & Pratt, T. (2004). Broadband MIMO-OFDM wireless communications, *Proceedings of the IEEE* 92(2): 271–294.

Tarokh, V., Jafarkhani, H. & Calderbank, A. (1999). Space-time block codes from orthogonal designs, *IEEE Transactions on Information Theory* 45(5): 1456–1467.

Telatar, E. (1999). Capacity of Multi-antenna Gaussian Channels, *European Transactions on Telecommunications* 10(6): 585–595.

van Zelst, A. & Hammerschmidt, J. (2002). A single coefficient spatial correlation model for multiple-input multiple-output (mimo) radio channels, *Proc. Proc. URSI XXVIIth General Assembly*, Maastricht (the Netherlands), pp. 1–4.

Vikalo, H. & Hassibi, B. (2005). On the sphere-decoding algorithm ii. generalizations, second-order statistics, and applications to communications, *Signal Processing, IEEE Transactions on* 53(8): 2819 – 2834.

Weinstein, S. & Ebert, P. (1971). Data transmission by frequency-division multiplexing using the discrete Fourier transform, *IEEE Trans. Commun. Tech.* 19: 628–634.

Wittneben, S. (1993). A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation, *Proc. IEEE Int. Conf. on Commun.*, Geneva (Switzerland), pp. 1630–1634.

Yee, N., Linnartz, J.-P. & Fettweis, G. (1993). Multi-carrier CDMA in indoor wireless radio networks, *Proc. IEEE Int. Symp. on Pers., Indoor and Mob. Rad. Comm.*, Yokohama (Japan), pp. 109–113.

6

Optimal Resource Allocation in OFDMA Broadcast Channels Using Dynamic Programming

Jesús Pérez, Javier Vía and Alfredo Nazábal
University of Cantabria
Spain

1. Introduction

OFDM (Orthogonal Frequency Division Multiplexing) is a well-known multicarrier modulation technique that allows high-rate data transmissions over multipath broadband wireless channels. By using OFDM, a high-rate data stream is split into a number of lower-rate streams that are simultaneously transmitted on different orthogonal subcarriers. Thus, the broadband channel is decomposed into a set of parallel frequency-flat subchannels; each one corresponding to an OFDM subcarrier. In a single user scenario, if the channel state is known at the transmitter, the system performance can be enhanced by adapting the power and data rates over each subcarrier. For example, the transmitter can allocate more transmit power and higher data rates to the subcarriers with better channels. By doing this, the total throughput can be significantly increased.

In a multiuser scenario, different subcarriers can be allocated to different users, which constitutes an orthogonal multiple access method known as OFDMA (Orthogonal Frequency Division Multiple Access). OFDMA is one of the principal multiple access schemes for broadband wireless multiuser systems. It has being proposed for use in several broadband multiuser wireless standards like IEEE 802.20 (MBWA: http://grouper.ieee.org/groups/802/20/), IEEE 802.16 (WiMAX: http://www.ieee802.org/16/, 2011) and 3GPP-LTE (http://www.3gpp.org/). This chapter focuses on the OFDMA broadcast channel (also known as downlink channel), since this is typically where high data rates and reliability is needed in broadband wireless multiuser systems. In OFDMA downlink transmission, each subchannel is assigned to one user at most, allowing simultaneous orthogonal transmission to several users. Once a subchannel is assigned to a user, the transmitter allocates a fraction of the total available power as well as a modulation and coding (data rate). If the channel state is known at the transmitter, the system performance can be significantly enhanced by allocating the available resources (subchannels, transmit power and data rates) intelligently according to the users' channels. The allocation of these resources determines the quality of service (QoS) provided by the system to each user. Since different users experience different channels, this scheme does not only exploit the frequency diversity of the channel, but also the inherent multiuser diversity of the system.

In multiuser transmission schemes, like OFDMA, the information-theoretic system performance is usually characterized by the capacity region. It is defined as the set of rates that can be simultaneously achieved for all users (Cover & Thomas, 1991). OFDMA is a suboptimal scheme in terms of capacity, but near capacity performance can be achieved when the system resources are optimally allocated. This fact, in addition to its orthogonality and feasibility, makes OFDMA one of the preferred schemes for practical systems. It is well known that coding across the subcarriers does not improve the capacity (Tse & Viswanath, 2005), so maximum performance is achieved by using separate codes for each subchannel. Then, the data rate received by each user is the sum of the data rates received from the assigned subchannels. The set of data rates received by all users for a given resource allocation gives rise to a point in the rate region. The points of the segment connecting two points associated with two different resource allocation strategies can always be achieved by time sharing between them. Therefore, the OFDMA rate region is the convex hull of the points achieved under all possible resource allocation strategies.

To numerically characterize the boundary of the rate region, a weight coefficient is assigned to each user. Then, since the rate region is convex, the boundary points are obtained by maximizing the weighted sum-rate for different weight values. In general, this leads to non-linear mixed constrained optimization problems quite difficult to solve. The constraint is given by the total available power, so it is always a continuous constraint. The optimization or decision variables are the user and the rate assigned to each subcarrier. The first is a discrete variable in the sense that it takes values from a finite set. At this point is important to distinguish between continuous or discrete rate adaptation. In the first case the optimization variable is assumed continuous whereas in the second case it is discrete and takes values from a finite set. The later is the case of practical systems where there is always a finite codebook, so only discrete rates can be transmitted through each subchannel. Unfortunately, regardless the nature of the decision variables, the resulting optimization problems are quite difficult to solve for realistic numbers of users and subcarriers.

This chapter analyzes the maximum performance attainable in broadcast OFDMA channels from the information-theoretic point of view. To do that, we use a novel approach to the resource allocation problems in OFDMA systems by viewing them as optimal control problems. In this framework the control variables are the resources to be assigned to each OFDM subchannel (power, rate and user). Once they are posed as optimal control problems, dynamic programming (DP) (Bertsekas, 2005) is used to obtain the optimal resource allocation. The application of DP leads to iterative algorithms for the computation of the optimal resource allocation. Both continuous and discrete rate allocation problems are addressed and several numerical examples are presented showing the maximum achievable performance of OFDMA in broadcast channels as function of different channel and system parameters.

1.1 Review of related works

Resource allocation in OFDMA systems has been an active area of research during the last years and a wide variety of techniques and algorithms have been proposed. The capacity region of general broadband channels was characterized in (Goldsmith & Effros, 2001), where the authors also derived the optimal power allocation achieving the boundary points of the capacity region. In this seminal work, the channel is decomposed into a set of N parallel independent narrowband subchannels. Each parallel subchannel is assigned to various users, to a single user, or even not assigned to any user. In the first case, the transmitter uses superposition coding (SC) and the corresponding receivers use successive interference cancelation (SIC). If a subchannel is assigned to a single user, an AWGN capacity-achieving code is used. Moreover, a fraction of the total available power is assigned to each user in each subchannel. Then, taking the limit as N goes to infinite (continuous frequency variable), the problem can be solved using multilevel water-filling. Similarly, in (Hoo et al., 2004) the authors characterize the asymptotic (when N goes to infinite) FDMA multiuser capacity region and propose optimal and suboptimal resource allocation algorithms to achieve the points in such region. Here, unlike (Goldsmith & Effros, 2001), each subchannel is assigned to one user at most and a separate AWGN capacity-achieving code is used in each subchannel.

In OFDMA systems the number of subchannels is finite. Each subchannel is assigned to one user at most, and a power value is allocated to each subcarrier. OFDMA is a suboptimal scheme in terms of capacity but, due to its orthogonality and feasibility, it is an adequate multiple access scheme for practical systems. Moreover, OFDMA can achieve near capacity performance when the system resources are optimally allocated.

In (Seong et al., 2006) and (Wong & Evans, 2008) efficient resource allocation algorithms are derived to characterize the capacity region of OFDMA downlink channels. The proposed algorithms are based on the dual decomposition method (Yu & Lui, 2006). In (Wong & Evans, 2008) the resource allocation problem is considered for both continuous and discrete rates, as well as for the case of partial channel knowledge at the transmitter. By using the dual decomposition method, the algorithms are asymptotically optimal when the number of subcarriers goes to infinite and is close to optimal for practical numbers of OFDM subcarriers.

Some specific points in the rate region are particularly interesting. For example the maximum sum-rate point where the sum of the users' rates is maximum, or the maximum symmetric-rates point where all users have maximum identical rate. Many times, in practical systems one is interested in the maximum achievable performance subject to various QoS (Quality of Service) users' requirements. For example, what is the maximum sum rate maintaining given proportional rates among users, or what is the maximum sum-rate guarantying minimum rate values to a subset of users. All these are specific points in the capacity region that can be achieved with specific resource allocation among the users. A crucial problem here is to determine the optimal resource allocation to achieve such points. Mathematically, these problems are also formulated as optimization problems constrained by the available system resources. In (Jang & Lee, 2003) the authors show the resource allocation strategy to maximize the sum rate of multiuser transmission in broadcast OFDM channels. They show that the maximum sum-rate is achieved when each subcarrier is assigned to the user with the best channel gain for that subcarrier. Then, the transmit power is distributed over the subcarriers by the water-filling policy. In asymmetric channels, the maximum sum-rate point is usually unfair because the resource allocation strategy favors users with good channel, producing quite different users' rates. Looking for fairness among users, (Ree & Cioffi, 2000) derive a resource allocation scheme to maximize the minimum of the users' rates. In (Shen et al., 2005) the objective is to maximize the rates maintaining proportional rates among users. In (Song & Li, 2005) an optimization framework based on utility-function is proposed to trade off fairness and efficiency. In (Tao et al., 2008), the authors maximize the sum rate guarantying fixed rates for a subset of users.

2. Channel and system model

Fig. 1 shows a block diagram of a single-user OFDM system with N subcarriers employing power and rate adaptation. It comprises three main elements: the transmitter, the receiver and the resource allocator. The channel is assumed to remain fixed during a block of OFDM symbols. At the beginning of each block the receiver estimates the channel state and sends this information (CSI: Channel state information) to the resource allocator, usually via a

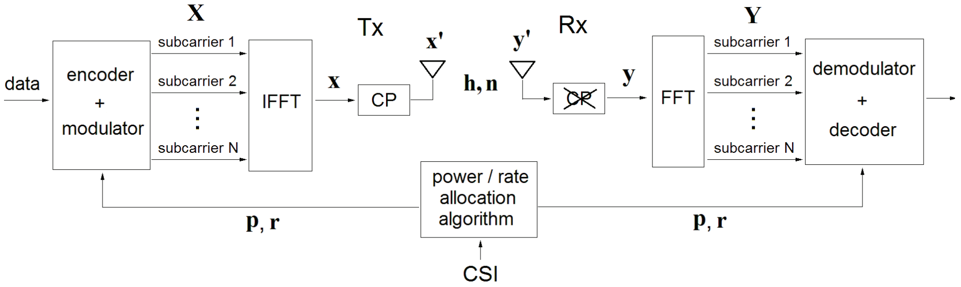


Fig. 1. Single-user OFDM system with power and rate adaptation.

feedback channel. The resource allocator can be physically embedded with the transmitter or the receiver. From the CSI, the resource allocation algorithm computes the data rate and transmit power to be transmitted through each subcarrier. Let vectors $\mathbf{r} = [r_1 r_2 \cdots r_N]^T$ and $\mathbf{p} = [p_1 p_2 \cdots p_N]^T$ denote the data rates and transmit powers allocated to the OFDM subchannels, respectively. This information is sent to the transmit encoder/modulator block, which encodes the input data according to $\mathbf{r}$ and $\mathbf{p}$, and produces the streams of encoded symbols to be transmitted through the different subchannels. It is well known that coding across the subcarriers does not improve the capacity (Tse & Viswanath, 2005) so, from a information-theoretic point of view, the maximum performance is achieved by using independent coding strategies for each OFDM subchannel. To generate an OFDM symbol, the transmitter picks one symbol from each subcarrier stream to form the symbols vector $\mathbf{X} = [X[1], X[2], \ldots, X[N]]^T$. Then, it performs an inverse fast Fourier transform (IFFT) operation on $\mathbf{X}$ yielding the vector $\mathbf{x}$. Finally the OFDM symbol $\mathbf{x'}$ is obtained by appending a cyclic prefix (CP) of length L_{cp} to $\mathbf{x}$. The receiver sees a vector of symbols $\mathbf{y'}$ that comprises the OFDM symbol convolved with the base-band equivalent discrete channel response $\mathbf{h}$ of length L, plus noise samples

$$\mathbf{y'} = \mathbf{h} * \mathbf{x'} + \mathbf{n}. \tag{1}$$

It is assumed that the noise samples at the receiver ($\mathbf{n}$) are realizations of a ZMCSCG (zero-mean circularly-symmetric complex Gaussian) random variables with variance σ^2: $\mathbf{n} \sim CN(\mathbf{0}, \sigma^2 \mathbf{I})$. The receiver strips off the CP and performs a fast Fourier transform (FFT) on the sequence $\mathbf{y}$ to yield $\mathbf{Y}$. If $L_{cp} \geq L$, it can be shown that

$$Y_k = H_k X_k + N_k, \quad k = 1, \ldots, N, \tag{2}$$

where $\mathbf{H} = [H_1, H_2, \ldots H_N]^T$ is the FFT of $\mathbf{h}$, i.e. the channel frequency response for each OFDM subcarrier, and the N_k's are samples of independent ZMCSCG variables with variance σ^2. Therefore, OFDM decomposes the broadband channel into N parallel subchannels with channel responses given by $\mathbf{H} = [H_1, H_2, \ldots H_N]^T$. In general the H_k's at different subcarriers are different.

Note that the energy of the symbol X_k is determined by the k-th entry of the power allocation vector p_k. It is assumed that the transmitter has a total available transmit power P_T to be distributed among the subcarriers, so $\sum_{k=1}^{N} p_k \leq P_T$. The coding/modulation employed for the k-th subchannel is determined by the corresponding entry (r_k) of the rate allocation vector $\mathbf{r}$.

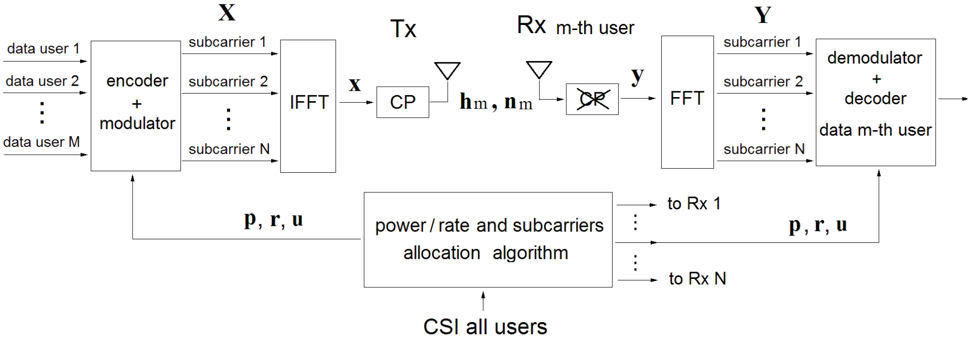


Fig. 2. Multi-user OFDM system with adaptive resources allocation.

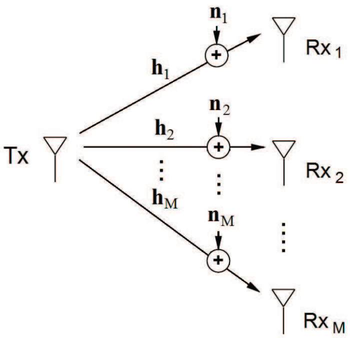


Fig. 3. M-users broadcast broadband channel.

Fig. 2 shows a block diagram of a downlink OFDMA system. It comprises the transmitter, the resource allocator unit and M users' receivers (Fig. 2 only shows the m-th receiver). The resource allocator is physically embedded with the transmitter. It is assumed that the transmitter sends independent information to each user. The base-band equivalent discrete channel response of the m-th user is denoted by $\mathbf{h}_m = [h_{m,1}h_{m,2}\cdots h_{m,L_m}]^T$, where now L_m is the number of channel taps and $\mathbf{n}_m \sim CN(\mathbf{0}, \sigma_m^2\mathbf{I})$ are the noise samples at the m-th receiver. Noise and channels at different receivers are assumed to be independent. A scheme of a M-user OFDM broadcast channel is depicted in Fig. 3.

Let $\mathbf{H}_m = [H_{m,1}H_{m,2}\cdots H_{m,N}]^T$ denote the complex-valued frequency-domain channel response of the OFDM channel, as seen by the m-th user, for the N subchannels. As it was mentioned, $\mathbf{H}_m$ is the N-points discrete-time Fourier transform (DFT) of $\mathbf{h}_m$.

It is assumed that the multi-user channel remains constant during the transmission of a block of OFDM symbols. At the beginning of each block each receiver estimates its channel response for each subcarrier, and informs the resource allocator by means of a feedback channel. Then, it computes the resource allocation vectors $\mathbf{r}$, $\mathbf{p}$ and $\mathbf{u} = [u_1u_2\ldots u_N]^T$, where u_k denotes the user assigned to the k-th subcarrier. Each subcarrier is assigned to a single user, so it is assumed that subcarriers are not shared by different users. Note that, since $u_k \in S_u = \{1, 2, \ldots M\}$, there are M^N possible values of $\mathbf{u}$, so M^N different ways to assign the subcarriers to the users. Once these vectors have been computed, the resource allocator informs the transmitter and receivers through control channels. Then, the transmitter encode the input data according to the resource allocation vectors and stores the stream of encoded symbols to be transmitted through the OFDM subchannels. The OFDM symbols are created and transmitted as in the single-user case. Each user receives and decodes its data from the assigned subchannels (given by **u**).

Let γ be a $M \times N$ matrix whose entries are the channel power gains for the different users and subcarriers normalized to the corresponding noise variance

$$\gamma_{m,k} = \frac{|H_{m,k}|^2}{\sigma_m^2}. \tag{3}$$

Assuming a continuous codebook available at the transmitter, r_k can take any value subject to the available power and the channel condition. The maximum attainable rate through the k-th subchannel is given by

$$r_k = \log_2(1 + p_k \gamma_{u_k,k}) \qquad \text{bits/OFDM symbol}, \tag{4}$$

where p_k is the power assigned to the k-th subchannel. The minimum needed power to support a given data rate r_k through the k-subcarrier will be

$$p_k = \frac{2^{r_k} - 1}{\gamma_{u_k,k}}. \tag{5}$$

We assume that the system always uses the minimum needed power to support a given rate so, for a fixed subcarriers-to-users allocation **u**, the r_k's and the p_k's are interchangeable in the sense that a given rate determines the needed transmit power and viceversa.

In practical systems there is always a finite codebook, so the data rate at each subchannel is constrained to take values from a discrete set $r_k \in S_r = \{r^{(1)}, r^{(2)}, \dots, r^{(N_r)}\}$ where each value corresponds to a specific modulation and code from the available codebook. The transmit rates and powers are related by

$$r_k = \log_2(1 + \beta(r_k) p_k \gamma_{u_k,k}) \qquad \text{bits/OFDM symbol}, \tag{6}$$

where the so-called SNR-gap approximation is adopted Cioffi et al. (1995), being $0 < \beta(r) \leq 1$ the SNR gap for the corresponding code (with rate r). For a given code $\beta(r)$ depends on a pre-fixed targeted maximum bit-error rate. Then, the SNR-gap can be interpreted as the penalty in terms of SNR due to the use of a realistic modulation/coding scheme. There will be a SNR gap $\beta(r^{(i)}), i = 1, \dots N_r$ associated with each code of the codebook for a given targeted bit-error rate. The minimum needed power to support r_k will be

$$p_k = \frac{2^{r_k} - 1}{\beta(r_k)\gamma_{u_k,k}}. \tag{7}$$

Since there is a finite number of available data rates, there will be a finite number of possible rate allocation vectors **r**. Note that there are $(N_r)^N$ possible values of **r**, but, in general, for a given **u** only some of them will fulfil the power constraint.

3. The rate region of OFDMA

For a given subcarriers-to-users and rates-to-subcarriers allocation vectors $\mathbf{u}$ and $\mathbf{r}$, the total rate received by the m-th user will be given by

$$R_m(\mathbf{r}, \mathbf{u}) = \sum_{k=1}^{N} \delta_{m,u_k} r_k \qquad \text{bits/OFDM symbol,} \tag{8}$$

where $\delta_{i,j}$ is the Kronecker delta. The users' rates are grouped in the corresponding rate vector

$$\mathbf{R}(\mathbf{r}, \mathbf{u}) = [R_1(\mathbf{r}, \mathbf{u}), R_2(\mathbf{r}, \mathbf{u}), \cdots, R_M(\mathbf{r}, \mathbf{u})]^T, \tag{9}$$

which is the point in the rate region associated with the resource allocation vectors $\mathbf{r}$ and $\mathbf{u}$. Let $\mathcal{R}_0$ denote the points achieved for all possible combinations of $\mathbf{u}$ and $\mathbf{r}$

$$\mathcal{R}_0 = \bigcup_{\mathbf{r} \in S_\mathbf{r}, \mathbf{u} \in S_\mathbf{u}} \mathbf{R}(\mathbf{r}, \mathbf{u}), \tag{10}$$

where $S_\mathbf{r}$ and $S_\mathbf{u}$ are the set of all possible rates-to-subcarriers and subcarriers-to-users allocation vectors, respectively. Therefore, $\mathcal{R}_0$ comprises the rate vectors associated with single resource allocation strategies given by $\mathbf{u}$ and $\mathbf{r}$. Later, it will be shown that, in general, $\mathcal{R}_0$ is not a convex region. Let $(\mathbf{r}_1, \mathbf{u}_1)$ and $(\mathbf{r}_2, \mathbf{u}_2)$ be two possible resource allocations that achieves the points $\mathbf{R}_1 = \mathbf{R}(\mathbf{r}_1, \mathbf{u}_1)$ and $\mathbf{R}_2 = \mathbf{R}(\mathbf{r}_2, \mathbf{u}_2)$ in $\mathcal{R}_0$. By time-sharing between the two resource allocation strategies, all points in the segment $\mathbf{R}_1$-$\mathbf{R}_2$ can be achieved. Therefore, the rate region of OFDMA will be the convex hull of $\mathcal{R}_0$: $\mathcal{R} = H(\mathcal{R}_0)$. Note that the achievement of any point of $\mathcal{R}$ not included in $\mathcal{R}_0$ requires time-sharing among different resource allocation schemes.

The next two subsections analyze the OFDMA rate region for the cases of continuous and discrete rates. Mathematical optimization problems for the computation of the rate region are posed, and their solution by means of the DP algorithm is presented.

3.1 Continuous rates

Let us first consider the achievable rate region $\mathcal{R}_0(\mathbf{u})$ for a fixed subcarriers-to-users allocation vector $\mathbf{u}$. It will be the union of the points achieved for all possible rates-to-subcarriers allocation vectors $\mathbf{r}$

$$\mathcal{R}_0(\mathbf{u}) = \bigcup_{\mathbf{r} \in S_\mathbf{r}} \mathbf{R}(\mathbf{r}, \mathbf{u}). \tag{11}$$

It can be shown that $\mathcal{R}_0(\mathbf{u})$ is a convex region (Cover & Thomas, 1991), being its boundary points the solution of the following convex optimization problems

$$\begin{aligned} \underset{\mathbf{r}}{\text{maximize}} \quad & \boldsymbol{\lambda}^T \mathbf{R}(\mathbf{u}) = \Sigma_{k=1}^{N} \lambda_{u_k} r_k \\ \text{subject to} \quad & \Sigma_{k=1}^{N} (2^{r_k} - 1)/\gamma_{u_k,k} \leq P_T \\ & r_k \in S_\mathbf{r}, \quad k = 1, \ldots, N, \end{aligned} \tag{12}$$

for different values of vector $\boldsymbol{\lambda} = [\lambda_1 \lambda_2 \cdots \lambda_M]^T$, where $\lambda_m \geq 0$. $\boldsymbol{\lambda}$ can be geometrically interpreted as the orthogonal vector to the hyperplane tangent to the achievable rate region at a point in the boundary. The components of $\boldsymbol{\lambda}$ are usually denoted as users' priorities. Note the constraint regarding the total available power. This is a well-known convex problem (Boyd & Vandenberghe, 2004) whose solution can be expressed in closed-form as follows

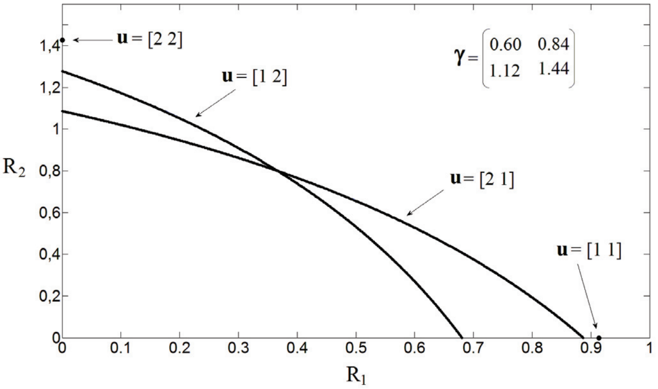


Fig. 4. Rate regions of all vectors **u**. The total transmit power is $P_T = 1$.

$$r_k^* = \begin{Bmatrix} \log_2\left(\frac{\lambda_{u_k}\gamma_{u_k,k}}{\mu}\right) & \text{if } \mu \le \lambda_{u_k}\gamma_{u_k,k} \\ 0 & \text{if } \mu \ge \lambda_{u_k}\gamma_{u_k,k} \end{Bmatrix} \tag{13}$$

where μ is a Lagrangian parameter which can be implicitly obtained from

$$\sum_{k=1}^{N}\left(\frac{\lambda_{u_k}}{\mu} - \frac{1}{\gamma_{u_k,k}}\right)^+ = P_T, \tag{14}$$

where $(a)^+ = \max\{a, 0\}$.

Fig. 4 shows the rate regions $\mathcal{R}(\mathbf{u})$ for all possible subcarriers-to-users allocation vectors $(\mathbf{u})$ in a toy example with $M = 2$ users, $N = 2$ subcarriers and $P_T = 1$. (Although it is not a realistic channel, it is used here to illustrate the resource allocation problem in OFDMA channels). Here and in the following results, the rates are given in bits/OFDM symbol.

When the system allocates all subcarriers to a single user ($u_k = u, \forall k$), the broadcast channel turns into a single-user channel and the solution of (12) does not depend on $\boldsymbol{\lambda}$. Therefore, in these cases the rate region degenerates in a single point on the corresponding axis. The rate at this point is the capacity of the corresponding single-user OFDM channel. Once the optimal rate vector $\mathbf{r}^*$ is obtained, the power to be allocated to each subcarrier is given by (5).

The achievable points for all possible values of **u** and **r** will be

$$\mathcal{R}_0 = \bigcup_{\mathbf{u}\in S_\mathbf{u}} \mathcal{R}_0(\mathbf{u}). \tag{15}$$

In general, $\mathcal{R}_0$ is not convex. This fact can be observed in the example of Fig. 4. The rate region of the OFDMA broadcast channel is the convex hull of $\mathcal{R}_0$: $\mathcal{R} = H(\mathcal{R}_0)$. The rate region for the example of Fig. 4 is depicted in Fig. 5 as the convex hull of the region achieved by all vectors **u**. In general, there are subcarriers-to-users allocation vectors **u** that are never optimal. This is the case of $u = [1, 2]^T$ in the example.

As it was mentioned, there are M^N possible subcarriers-to-users allocation vectors **u**. Therefore, an exhaustive search among all the possible vectors requires the computation of

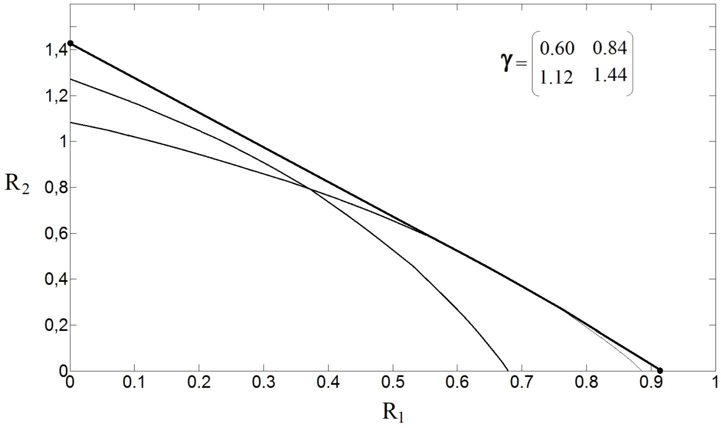


Fig. 5. OFDMA rate region. It is the convex hull of the rate regions achieved for all vectors **u** (see Fig. 4).

M^N waterfilling solutions for each vector $\boldsymbol{\lambda}$, which is not feasible for practical values of N and M. An alternative is to jointly optimize over **u** and **p** simultaneously, so the problem becomes

$$\begin{array}{ll} \underset{\mathbf{r},\mathbf{u}}{\text{maximize}} & \boldsymbol{\lambda}^T\mathbf{R}(\mathbf{r},\mathbf{u}) = \Sigma_{k=1}^{N}\lambda_{u_k} r_k \\ \text{subject to} & \Sigma_{k=1}^{N}(2^{r_k}-1)/\gamma_{u_k,k} \le P_T \\ & u_k \in S_{\mathbf{u}}, \quad k = 1,\dots,N \\ & r_k \in S_{\mathbf{r}}, \quad k = 1,\dots,N \end{array} \tag{16}$$

This is a mixed non-linear constrained optimization problem. In general these kind of problems are difficult to solve. However, it has the structure of a DP problem with the following elements (see appendix: Dynamic Programming):

- The process stages are the subchannels, so the number of stages is N,
- Control vector: $\mathbf{c}_k = [u_k, r_k]^T$,
- State variable: $x_k = \Sigma_{i=1}^{k-1}(2^{r_i}-1)/\gamma_{u_i,i}$,
- Initial state $x_1 = 0$,
- Subsets of possible states: $0 \le x_k \le P_T$,
- Subsets of admissible controls:

$$C_k(x_k) = \left\{ [u_k, r_k]^T \;\middle|\; u_k \in S_{\mathbf{u}}, \;\; r_k \in S_{\mathbf{r}}, r_k \le \log_2(1 + (P_T - x_k)\gamma_{u_k,k}) \right\},$$

- System equation: $x_{k+1} = f_k(x_k, c_k) = x_k + (2^{r_k}-1)/\gamma_{u_k,k}$,
- Cost functions: $g_k(c_k) = \lambda_{u_k} r_k$.

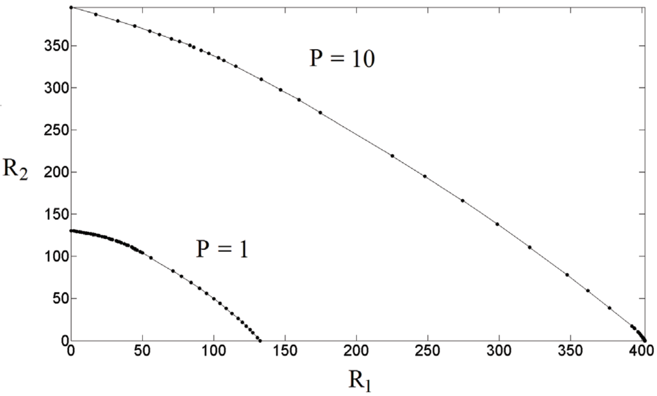


Fig. 6. Rate regions for the two-users channel of Fig. 7 considering different values of average transmit power per subchannel (P).

The entries of the control vector $\mathbf{c}_k$ are the user and the rate allocated to the k-th subchannel, that take values from the sets $S_{\mathbf{u}}$ and $S_{\mathbf{r}}$, respectively. The state variable x_k is the accumulated power transmitted in the previous subchannels. Therefore $0 \leq x_k \leq P_T$, and the initial state is $x_1 = 0$. The control component r_k is constrained by the available power at the k-th stage: $P_T - x_k$. Note that the solutions of (16) for different $\boldsymbol{\lambda}$'s are the points of $\mathcal{R}_0$ located in the boundary of the rate region, and the convex hull of these points are the boundary of the rate region.

By using the DP algorithm, rate regions for a more realistic two user channel have been computed. They are depicted in Fig. 6. In this example the number of subcarriers is $N = 128$ and the users' subchannel responses are shown in Fig. 7. They are normalized so the average gain (averaging over the subchannels and users) equals 1. These channel realizations have been obtained from a broadband Rayleigh channel model with $L = 16$ taps and an exponential power delay profile with decay factor $\rho = 0.4$. This channel model will be described in section 5. The noise variances are assumed to be $\sigma_m^2 = 1$, identical for all users. Fig. 6 shows the rate region for two different values of average power per subchannel: $P = P_T/N = 1$ and $P = P_T/N = 10$. Note that if the OFDM subchannels were identical (frequency-flat broadband channel), the average SNR at the OFDM subchannels would be 0dB and 10dB, respectively.

To obtain the rate regions of Fig.6, (16) has to be solved for each $\boldsymbol{\lambda}$ by using the DP algorithm. In each case, the solution is a pair of optimal resource allocation vectors $\mathbf{u}^*$ and $\mathbf{r}^*$. Then, the power to be transmitted through the subchannels $\mathbf{p}^*$ is given by (5). The corresponding users' rate vector are obtained from (8) and shown in Fig. 6 as a marker point in the boundary of the rate region. Therefore, the marker points are the points of $\mathcal{R}_0$ located in the boundary of the rate region and associated with pairs of resource allocation vectors $(\mathbf{u}, \mathbf{r})$ which are solutions of (16) for different users' priority vectors. For example, for $\boldsymbol{\lambda} = [0.4, 0.6]^T$ and $P_T = 10$ the optimal resource allocation vectors are shown in Fig. 8, as well as the transmit power through the OFDM subchannels ($\mathbf{p}^*$). The resulting users' rates are $\mathbf{R}(\mathbf{u}^*, \mathbf{p}^*) = [82.7, 350.7]^T$.

Note that different vectors $\boldsymbol{\lambda}$ can lead to identical solution of (16), and hence to identical points/markers in the boundary of the rate region. The convex hull of the marker points

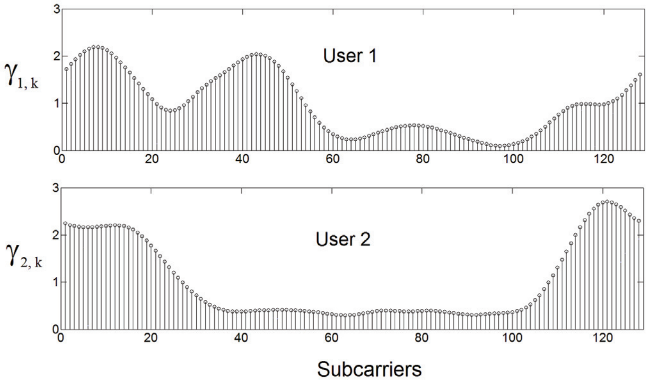


Fig. 7. Normalized subchannel gains at the OFDM subchannels.

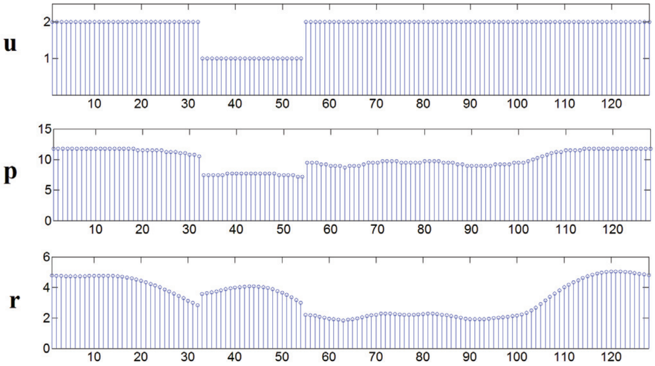


Fig. 8. Optimal resource allocation vectors **u** and **r** for $\boldsymbol{\lambda} = [0.4, 0.6]^T$ and $P_T = 10N$. The figure also shows the resulting power allocation vector **p**.

constitutes the boundary of the rate region. Any point in the segments between two markers is achieved by time sharing between the corresponding optimal resource allocations.

The application of the DP algorithm requires the control and state spaces to be discrete. Therefore, if they are continuous, they must be discretized by replacing the continuous spaces by discrete ones. Once the discretization is done, the DP algorithm is executed to yield the optimal control sequence for the discrete approximating problem. Hence, it becomes necessary to study the effect of discretization on the optimality of the solution. In the problem (16), the state variable x_k and the second component of the control vetctor (r_k) are continuous. To obtain the rate regions of Fig. 6, $S_{\mathbf{r}}$ was uniformly discretized considering $N_d = 2000$ possible rate values between 0 and a maximum rate which is achieved when the total power P_T is assigned to the best subchannel of all users. The state variable x_k was discretized in N_d possible values accordingly. To study the effect of discretization of x_k and r_k the rate regions for different values of N_d are shown in Fig. 9. The channels and simulation parameters were as in Fig. 6. It shows that the required N_d is less than 500.

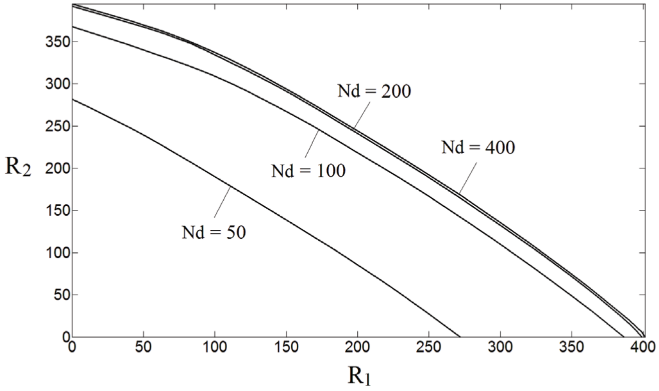


Fig. 9. Rate regions for different number of rate discretization values N_d

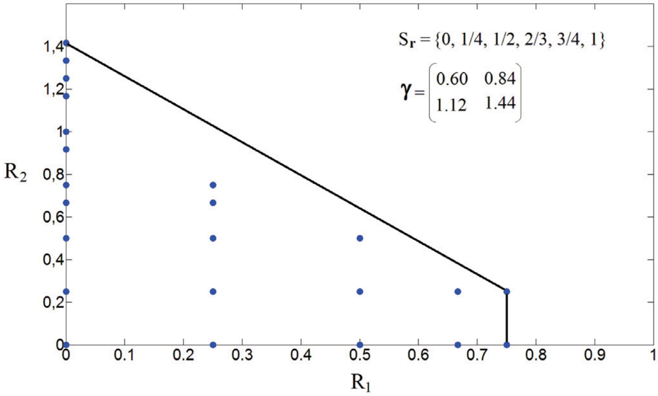


Fig. 10. Example of OFDMA rate region with discrete codebook $S_{\mathbf{r}}$.

3.2 Discrete rates

Now, $S_{\mathbf{r}}$ is a finite set and therefore the set of achievable points $\mathcal{R}_0$ is finite. It comprises all points resulting from the combinations of **r** and **u** that fulfill the power constraint. Therefore, the cardinality of $\mathcal{R}_0$ depends on γ, $S_{\mathbf{r}}$ and P_T.

As an example, let us consider again the channel example of Fig. 4 and 5 with $P_T = 1$ and a codebook with the following available rates $S_{\mathbf{r}} = \{0, 1/4, 1/2, 2/3, 3/4, 1\}$. Note that by including zero rate in S_r we consider the possibility of no transmission trough some subchannels. All achievable rate vectors ($\mathcal{R}_0$), and their convex hull, are shown in Fig. 10, where $\beta(r) = 1, \forall r \in S_{\mathbf{r}}$ is assumed.

The set $\mathcal{R}_0$ can be viewed as the union of the points achieved by different vectors $\mathbf{u} \in S_u$: $\mathcal{R}_0(\mathbf{u})$. For example, in Fig. 11 the points achieved by $\mathbf{u} = [1, 2]^T$ are highlighted, as well as their convex hull. It can be observed that, for this particular value of **u** the attainable rate vectors are always under the convex hull of $\mathcal{R}_0$. Therefore, $\mathbf{u} = [1, 2]$ is not the optimal subcarriers-to-users allocation vector in any case.

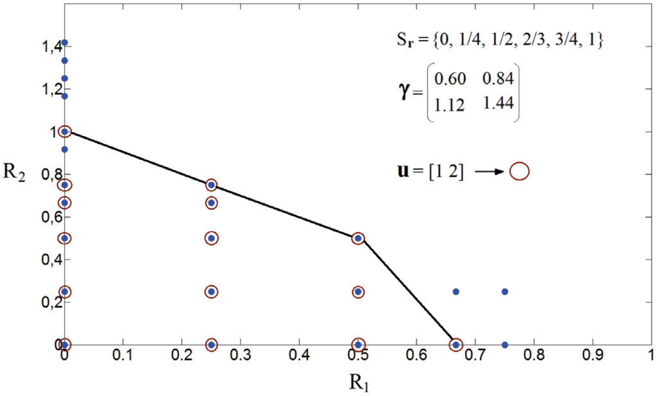


Fig. 11. Achievable points for $\mathbf{u} = [1, 2]$, and their convex hull.

In general the vertex points in the boundary of $\mathcal{R}_0(\mathbf{u})$ will be the solutions of (12), but now $S_\mathbf{r}$ is a finite set. Therefore, both the state and control variables are discrete so (12) is a fully integer optimization problem. Unlike the continuous rates case, there is not closed-form solution when the allowed rates belong to a discrete set. The problem can be formulated as a DP problem where

- The process stages are the subchannels, so the number of stages is N,
- Control variable: $\mathbf{c}_k = r_k$,
- State variable: $x_k = \Sigma_{i=1}^{k-1}(2^{r_i} - 1)/\gamma_{u_i,i}$, $0 \leq x_k \leq P_T$, $x_1 = 0$, $x_{N+1} = P_T$
- Initial state $x_1 = 0$
- Subsets of possible states: $0 \leq x_k \leq P_T$
- Subsets of admissible controls: $C_k(x_k) = \{r_k | r_k \in S_\mathbf{r}, r_k \leq \log_2(1 + (P_T - x_k)\gamma_{u_k,k})\}$,
- System equation: $x_{k+1} = f_k(x_k, c_k) = x_k + (2^{r_k} - 1)/\gamma_{u_k,k}$,
- Cost functions: $g_k(c_k) = \lambda_{u_k} r_k$.

The control variables $\mathbf{c}_k$ are the rates allocated to the subchannels, that take values from the set $S_\mathbf{r}$. The state variable x_k is the accumulated power transmitted in the previous subchannels (up to $k-1$-th subchannel). Therefore $0 \leq x_k \leq P_T$, and the initial state is $x_1 = 0$. The control component r_k is constrained by the available power at the k-th stage: $P_T - x_k$.

Since there are M^N possible values of $\mathbf{u}$, an exhaustive search among all possible vectors $\mathbf{u}$ requires to solve the above DP problem M^N times for each vector $\boldsymbol{\lambda}$, which is not feasible for practical values of M and N. In this case one has to jointly optimize over $\mathbf{u}$ and $\mathbf{r}$ simultaneously, as in (16). Now, unlike the continuous rates case, (16) is an integer programming problem because the control variable c_k is fully discrete taking values from a finite set $S_\mathbf{u} \times S_\mathbf{r}$.

Fig. 12 shows the rate regions for the two user channel of Fig. 7 considering continuous and discrete rate allocation. As it is expected, continuous rate adaptation always outperforms the discrete rate case. In all cases the rate region has been obtained from the DP algorithm. The figure depicts the rate region for two different values of average power per subcarrier: $P = 1$ and $P = 10$. It is also assumed that the noise at the OFDM subchannels are

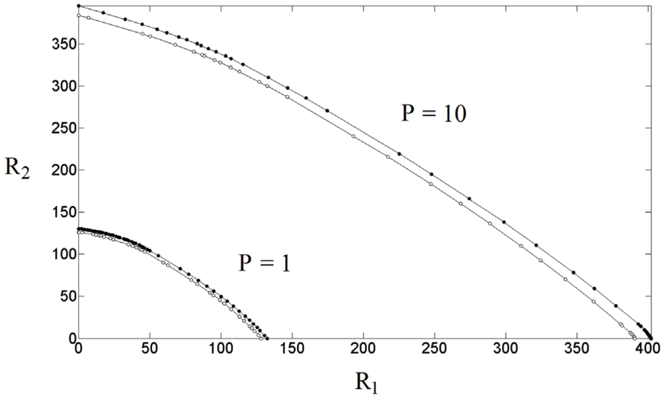


Fig. 12. Rate regions for the two user channel of Fig. 7 considering continuous and discrete rate allocation. P denotes the average transmit power per subchannel

i.i.d. with variance $\sigma_m^2 = 1$, identical for all users. Then, if the subchannels were identical and frequency-flat, the average SNR at the OFDM subchannels would be 0dB and 10dB, respectively. Now, the following set of available rates have been considered: $S_{\mathbf{r}} = \{0, 1/2, 3/4, 1, 3/2, 2, 3, 4, 9/2\}$. These are the data rates of a set of rate-compatible punctured convolutional (RCPC) codes, combined with M-QAM modulations, that are used in the 802.11a (http://grouper.ieee.org/groups/802/11/). The advantage of RCPC codes is to have a single encoder and decoder whose error correction capabilities can be modified by not transmitting certain coded bits (puncturing). Therefore, the same encoder and decoder are used for all codes of the RCPC codebook. This makes the RCPC codes, combined with adaptive modulation, a feasible rate adaptation scheme in wireless communications. Apart from the 802.11, punctured codes are used in other standards like WIMAX (http://www.ieee802.org/16/, 2011).

To obtain the rate regions of Fig. 12, the maximization problem (16) has been solved, using the DP algorithm, for each $\boldsymbol{\lambda}$. In each case, the solution is a pair of optimal resource allocation vectors $\mathbf{u}^*$ and $\mathbf{r}^*$. Now, the corresponding users' rate vector is obtained from (8) and shown in 12 as a marker point in the boundary of the rate region. For example, for $\boldsymbol{\lambda} = [0.4, 0.6]^T$ and $P_T = 10$ the optimal resource allocation vectors are shown in Fig. 13, as well as the transmit power assigned to each OFDM subchannel ($\mathbf{p}$). These vectors produces the users' rate vector $\mathbf{R}(\mathbf{u}^*, \mathbf{r}^*) = [89.5, 340.0]^T$. Note that this is quite similar to the users' rate vector attained for this $\boldsymbol{\lambda}$ in the continuous rate case. Comparing 8 and 13 one can observe that the resource allocation vectors are similar in the cases of continuous and discrete rates. In fact, for this particular case, the subcarriers-to-users allocation vectors $\mathbf{u}$ are identical and the rates-to-subcarriers allocation vectors $\mathbf{r}$ are quite similar. Similar behavior is observed for any other vector $\boldsymbol{\lambda}$.

4. Maximum sum-rate

In the case of continuous rate adaptation, the maximum sum rate will be the solution of (16) for $\boldsymbol{\lambda} = \mathbf{1}_M$. But, it was shown in (Jang & Lee, 2003) that the sum-rate is maximized when each subcarrier is assigned to the user with the best channel

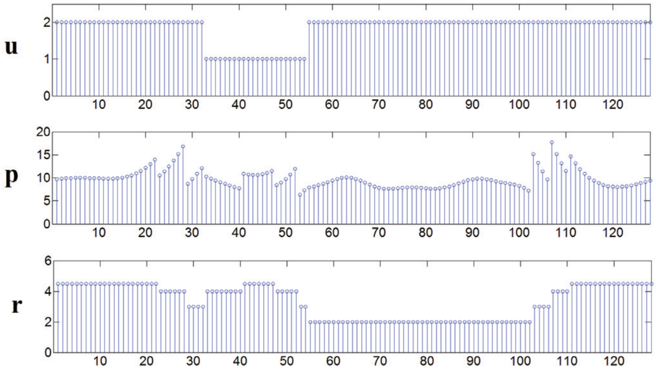


Fig. 13. Optimal resource allocation vectors **u** and **r** for $\boldsymbol{\lambda} = [0.4, 0.6]^T$ and $P_T = 10N$. The figure also shows the power transmitted through the OFDM subchannels **p**

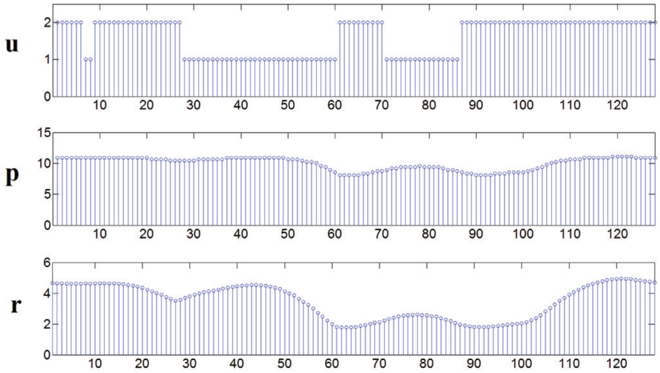


Fig. 14. Optimal resource allocation vectors **u** and **r** to achieve the maximum sum-rate. The total transmit power is $P_T = 10N$. The figure also shows the power allocation vector **p**

$$u_k^* = \arg\max_m \{\gamma_{m,k}\}, \quad k = 1, \ldots, N, \tag{17}$$

Then, the optimal rates can be calculated from (13) and (14) with $\boldsymbol{\lambda} = \mathbf{1}_M$, and the power allocated to each subcarrier is given by (5).

The maximum sum-rate point is always in the boundary of the rate region. In the channel of Figs. 4 and 5, the maximum sum rate is 1.43, which is achieved by $\mathbf{u}^* = [2, 2]^T$ and $\mathbf{r}^* = [0.53, 0.90]^T$. The power allocation is $\mathbf{p}^* = [0.40, 0.60]^T$. In this particular case the maximum sum-rate is achieved by allocating all the system resources to the user 2, so the rate for user 1 is zero. In the two-users channel of Fig. 7, the maximum sum rate, for $P = 10$, is 445.64. This is achieved by the resource allocation vectors depicted in Fig. 14.

In the case of discrete rate adaptation, the maximum sum rate is also achieved by allocating each subcarrier to the user with the best channel on it. But now, the rates allocated to the subchannels will be the solution of (12) with **u** given by (17). So, the optimal rate allocation

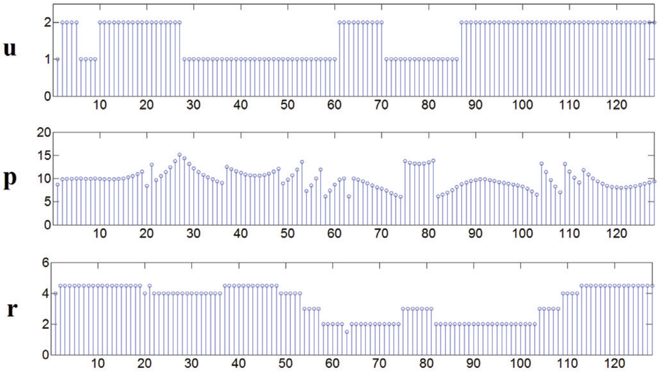


Fig. 15. Optimal resource allocation vectors **u** and **r** to achieve the maximum sum-rate. The total transmit power is $P_T = 10N$. The figure also shows the power allocation vector **p**

can also be obtained from the DP algorithm. In the simple case of Fig. 10, the maximum sum rate is 1.417, which is achieved by $\mathbf{u}^* = [2, 2]^T$, $\mathbf{r}^* = [2/3, 3/4]^T$ and $\mathbf{p}^* = [0.52, 0.47]^T$. Like in continuous rate adaptation, the maximum sum-rate point is always in the boundary of the rate region. In the two-user channel of Fig. 7, the maximum sum rate is 441.0 assuming that the average transmit power per subcarrier is $P = 10$ and the set of available rates is $S_{\mathbf{r}} = \{0, 1/2, 3/4, 1, 3/2, 2, 3, 4, 9/2\}$. It is achieved by the resource allocation vectors depicted in Fig. 15. Again, the maximum sum-rate and the corresponding resource allocation vectors are quite similar to the continuous rate case.

5. Outage rate region

The previous results show the achievable performance (rate vectors) for specific channel realizations. However, due to the intrinsic randomness of the wireless channel, the channel realizations can be quite different. To study the performance for all channel conditions we resort to the outage rate region concept Lee & Goldsmith (2001). The outage rate region for a given outage probability P_{out} consists of all rate vectors $\mathbf{R} = [R_1, R_2, \ldots, R_M]^T$ which can be maintained with an outage probability no larger than P_{out}. Therefore, the outage rate region will depend on the statistical parameters of the broadband channel.

In the following results the so-called broadband Rayleigh channel model is considered. This a widely accepted model for propagation environments where there is not line of sight between the transmitter and receiver. According to this model the time-domain channel response for the m-th user $\mathbf{h}_m$ is modeled as an independent zero-mean complex Gaussian random vector $\mathbf{h}_m \sim CN(\mathbf{0}, \text{diag}(\mathbf{\Gamma}_m))$, where $\mathbf{\Gamma}_m = [\Gamma_{m,1}\Gamma_{m,2}\ldots\Gamma_{m,L}]^T$ is the channel power delay profile (PDP), which is assumed to decay exponentially

$$\Gamma_{m,l} = E\{h_{m,l}h^*_{m,l}\} = A_m\rho^l_m, \quad l = 1, \ldots, L_m, \tag{18}$$

where L_m is the length of $\mathbf{h}_m$, ρ_m is the exponential decay factor and A_m is a normalization factor given by

$$A_m = E_m \frac{1 - \rho_m}{\rho_m(1 - \rho_m^{L_m})}, \tag{19}$$

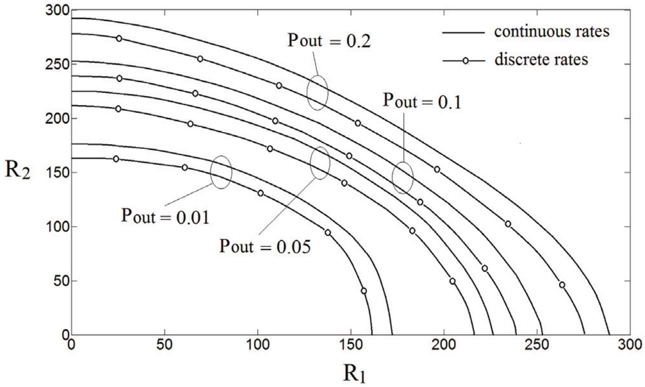


Fig. 16. Outage rate regions for different values of probability of outage P_{out}.

being E_m the average energy of the m-th user channel. Note that the frequency selectivity of the channel is determined by ρ_m, so the higher the ρ_m the higher is the frequency selectivity of the m-th user channel. The exponential decay PDP model is a widely used and it will be assumed in the following results. Any other PDP model could be used. Unless otherwise indicated, the parameters of the following simulations are

- Number of OFDM subcarriers: N=128
- i.i.d. Rayleigh fading channel model with $\rho = 0.4$, $L = 16$ and $E = 1$, for all users
- Available transmit power: $P_T = 10N$
- Same probability of outage for all users: $P_{out} = 0.1$
- In the case of discrete rates, $S_{\mathbf{r}} = \{0, 1/2, 3/4, 1, 3/2, 2, 3, 4, 9/2\}$

To obtain the outage rate region, 5000 channel realizations have been considered in each case. The rate region of each channel realization has been obtained by solving (16) with the DP algorithm.

Fig. 16 shows the outage rate regions for different values of outage probability (P_{out}). One can observe the performance gap between continuous and discrete rates, which is nearly constant for different values of P_{out}. Since the channel is identically distributed for both users, the rate regions are symmetric.

Fig. 17 shows the outage rate regions when the users' channels have different average energy (E_m). The sum of the average energy of the channels equals the number of users (2). In this case, only continuous rate adaptation is considered. As it is expected, the user with the best channel gets higher rates.

Fig. 18 shows the two-user outage rate regions for different values of average transmit power per subcarrier $P = P_T/N$. Note that the performance gap between continuous and discrete rates increases with P.

Finally, Fig. 19 compares the outage rate regions for different values of channel frequency selectivity. The figure clearly shows that the higher the frequency selectivity the more useful is the resource adaptation. The gap between continuous and discrete rates does not depend on the frequency selectivity of the channel.

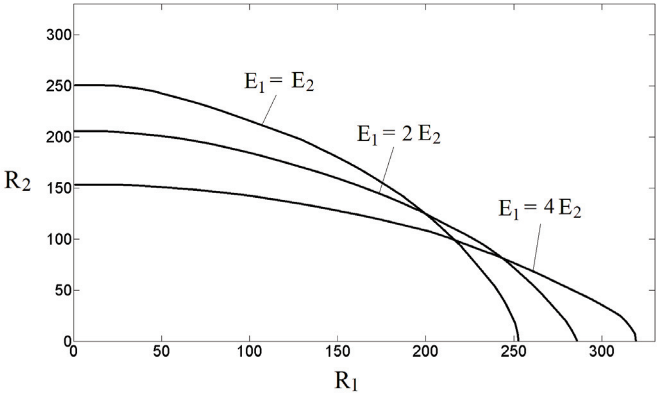


Fig. 17. Outage rate regions when the users' channels have different average energy (E_m).

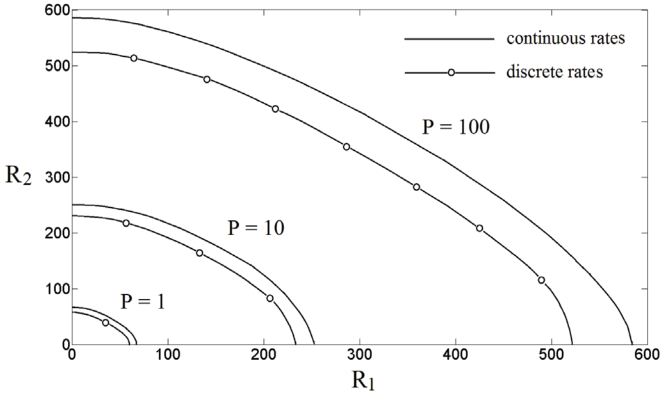


Fig. 18. Outage rate regions for different values of average power per subcarrier P.

6. Conclusions

This chapter analyzes the attainable performance of OFDMA in broadband broadcast channels from an information-theoretic point of view. Assuming channel knowledge at the transmitter, the system performance is maximized by optimally allocating the available system resource among the users. The transmitter has to assign a user, a fraction of the available power and a data rate (modulation and channel coding) to each subchannel. The optimal allocation of these resources leads to non-linear constrained optimization problems which, in general, are quite difficult to solve. These problems are solved by means of a novel approach to the resource allocation problems in OFDMA systems by viewing them as optimal control problems, where the control variables are the resources to be allocated to the OFDM subchannels (power, rate and user). Once the problems are posed as optimal control problems, dynamic programming is used to obtain the optimal resource allocation that maximizes the system performance. This constitutes a new methodology for the computation of optimal resource allocation in OFDMA systems. The achievable performance for given

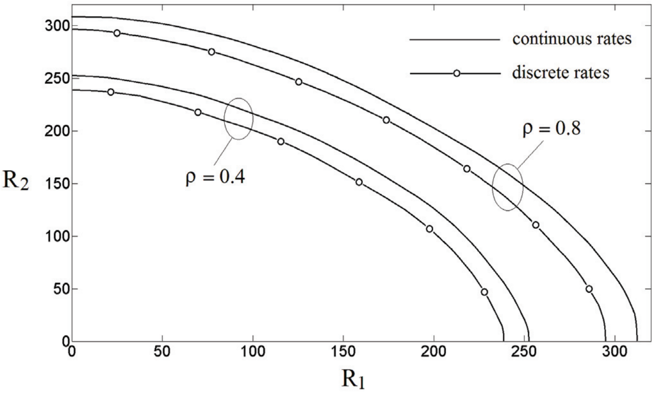


Fig. 19. Outage rate regions for different values of PDP exponential decay factor (ρ).

channel realization is characterized by the rate region, whereas the overall performance of OFDMA in random channels is characterized by means of the outage rate region. The cases of continuous and discrete rate allocation are addressed under a general framework. The first case leads to mixed optimization problems, whereas in the second case, the optimal resource allocation is the solution of integer optimization problems.

By using the dynamic programming algorithm, the achievable rate region of OFDMA for different channels and system parameters has been computed. The simulation results shows that the performance gap between continuous and discrete rate adaptation is quite narrow when the average transmit power per subcarrier is low, and it increases for higher values of transmit power. The frequency selectivity of the broadband channel has a important influence in the system performance. The higher the frequency selectivity, the more useful is the adequate resource adaptation. The gap between continuous and discrete rate adaptation does not depend on the frequency selectivity of the channel. The gap between continuous and discrete rate adaptation remains nearly constant for different values of outage probability.

7. Appendix: Dynamic programming

Dynamic Programming (DP) is a well-known mathematical technique for solving sequential control optimization problems with accumulative cost functions Bertsekas (2005). In a DP problem there is always an underlying dynamical discrete process governed by a set of system functions with the form

$$\mathbf{x}_{k+1} = f_k(\mathbf{x}_k, \mathbf{c}_k), \qquad 1 \leq k \leq N \tag{20}$$

where k is an index denoting a stage in the system evolution, N is the number of stages, $\mathbf{x}_k$ represents the state of the system at stage k and $\mathbf{c}_k$ denotes the control action to be selected at stage k (see Fig. 20). In general $\mathbf{x}_k \in S_k \subset R^n$, so the subset of possible states (S_k) can be different at different stages. In fact there is a fixed initial state $\mathbf{x}_1 = \mathbf{x}^{(0)}$, so the set of all possible states at the first stage has an unique value $S_1 = \{\mathbf{x}^{(0)}\}$. The control vector $\mathbf{c}_k$ at each stage are constrained to take values in a subset $C_k(\mathbf{x}_k)$, which, in general, depends on the current state $\mathbf{x}_k$ and on the stage (k).

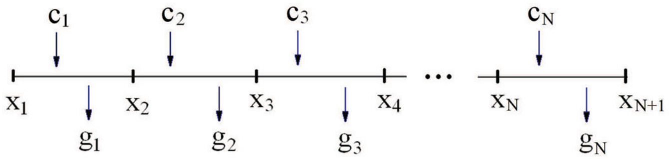


Fig. 20. Scheme of a DP problem showing the system states ($\mathbf{x}_k$), controls ($\mathbf{c}_k$) and costs (g_k) at different stages

At each state the system incurs in an additive cost $g_k(\mathbf{x}_k, \mathbf{c}_k)$, which, in general, depends on the state and on the applied control. The objective is to minimize the total cost of the system along its evolution by selecting the optimal controls at each stage

$$\min_{\{\mathbf{c}_k \in C_k(\mathbf{x}_k)\}_{k=1}^{N}} \quad \sum_{k=1}^{N} g_k(\mathbf{x}_k, \mathbf{c}_k). \tag{21}$$

Therefore, the elements of a DP problem are:

- The number of stages N,
- The control vector $\mathbf{c}_k$,
- The state vector: x_k
- The initial state: x_1
- The subset of possible states at each stage S_k
- The subset of possible controls at each stage $C_k(x_k)$,
- System equation: $x_{k+1} = f_k(x_k, c_k)$,
- Cost functions: $g_k(c_k)$.

In general, these problems are difficult to solve. A key aspect is that controls cannot be viewed in isolation since the controller must balance the cost at the current stage with the costs at future stages. The DP algorithm captures this trade-off. The DP algorithm simplifies (21) by breaking it down into simpler subproblems in a backwardly recursive manner Bellman (1957), which is described in the following lines.
The optimal cost from state $\mathbf{x}_k$ at stage k can be expressed as follows

$$\begin{aligned} J_k^*(\mathbf{x}_k) &= \min_{\mathbf{c}_k \in C_k(\mathbf{x}_k)} \quad \left\{ g_k(\mathbf{x}_k, \mathbf{c}_k) + J_{k+1}^*(f_k(\mathbf{x}_k, \mathbf{c}_k)) \right\}, \quad k = N-1, \ldots, 1 \\ J_N^*(\mathbf{x}_N) &= \min_{\mathbf{c}_N \in C_k(\mathbf{x}_N)} \quad g_k(\mathbf{x}_N, \mathbf{c}_N), \end{aligned} \tag{22}$$

and the optimal control policy at stage k for state $\mathbf{x}_k$ is

$$\boldsymbol{\mu}_k^*(\mathbf{x}_k) = \arg \min_{\mathbf{c}_k \in C_k(\mathbf{x}_k)} \quad \left\{ g_k(\mathbf{x}_k, \mathbf{c}_k) + J_{k+1}^*(f_k(\mathbf{x}_k, \mathbf{c}_k)) \right\}, \quad k = N, \ldots, 1$$

Finally, the optimal control sequence $\mathbf{c}^* = [\mathbf{c}_1^*, \mathbf{c}_2^*, \ldots, \mathbf{c}_N^*]$ and the corresponding system evolution $\mathbf{x}^* = [\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_{N+1}^*]$ are easily obtained by iteratively applying the optimal control policies from the initial state as follows

$$\begin{aligned} &\mathbf{x}_1^* = \mathbf{x}^{(0)}, \\ &\mathbf{c}_k^* = \boldsymbol{\mu}_k^*(\mathbf{x}_k), \quad \mathbf{x}_{k+1}^* = f_k(\mathbf{x}_k^*, \mathbf{c}_k^*), \quad k = 1, \dots N \end{aligned} \tag{23}$$

8. Acknowledgements

This work was supported by the Spanish Government, Ministerio de Ciencia e Innovacit'on (MICINN), under project COSIMA (TEC2010-19545-C04-03) and COMONSENS (CSD2008-00010, CONSOLIDER-INGENIO 2010 Program).

9. References

Bellman, R. E. (1957). *Dynamic Programming*, Princeton University Press, Princeton, NJ, USA.

Bertsekas, D. P. (2005). *Dynamic programming and optimal control, volume I*, Athena Scientific, Belmont, Massachusetts, USA.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press, Cambridge, UK.

Cioffi, J., Dudevoir, G., Eyuboglu, M. & Forney Jr, G. (1995). Mmse decision-feedback equalizers and coding ii. coding results, *IEEE Transactions on Communications* 43(10): 2595–2604.

Cover, T. M. & Thomas, J. A. (1991). *Elements of information theory*, Wiley.

Goldsmith, A. & Effros, M. (2001). The capacity region of broadcast channels with intersymbol interference and colored gaussian noise, *IEEE Transactions on information Theory* 47(1): 219–240.

Hoo, L., Halder, B., Tellado, J. & Cioffi, J. (2004). Multiuser transmit optimization for multicarrier broadcast channels: Asymptotic fdma capacity region and algorithms, *IEEE Transactions on Communications* 52(6): 922–930.

Jang, J. & Lee, K. (2003). Transmit power adaptation for multiuser ofdm systems, *IEEE Journal of Selected Areas in Communications* 21(2): 171–178.

Lee, L. & Goldsmith, A. (2001). Capacity and optimal resource allocation for fading broadcast channels - part ii: Outage capacity, *IEEE Transactions on information Theory* pp. 1103–1127.

Ree, W. & Cioffi, J. (2000). Increase in capacity of multiuser ofdm system using dynamic subchannel allocation, *Proceedings of IEEE Vehicular Technology Conference (VTC'00)*, Tokyo, Japan, pp. 1085–1089.

Seong, K., Mohseni, M. & Cioffi, J. M. (2006). Optimal resource allocation for ofdma downlink systems, *Proceedings of 2006 IEEE International Symposium on Information Theory (ISIT'06)*, Seattle, Washington, USA.

Shen, Z., Andrews, J. G. & Evans, B. L. (2005). Adaptive resource allocation in multiuser ofdm systems with proportional fairness, *IEEE Transactions on Wireless Communications* 4(6): 2726–2737.

Song, G. & Li, Y. (2005). Cross-layer optimization of ofdm wireless networks - part ii: Algorithm development, *IEEE Transactions on Wireless Communications* 4(2): 625–634.

Tao, M., Liang, Y. C. & Zhang, F. (2008). Resource allocation for delay differentiated traffic in multiuser ofdm systems, *IEEE Transactions on Wireless Communications* 7(6): 2190–2201.

Tse, D. & Viswanath, P. (2005). *Fundamentals of Wireless Communications*, Cambridge University Press, Cambridge, UK.

Wong, I. & Evans, B. (2008). *Resource Allocation in Multiuser Multicarrier Wireless Systems*, Springer, New York, USA.

Yu, W. & Lui, R. (2006). Dual methods for nonconvex spectrum optimization of multicarrier systems, *IEEE Transactions on Communications* 54(7): 1310–1322.

Primary User Detection in Multi-Antenna Cognitive Radio

Oscar Filio[1], Serguei Primak[1] and Valeri Kontorovich[2]
[1]*The University of Western Ontario*
[2]*Centre of Research and Advanced Studies (CINVESTAV-IPN)*
[1]*Canada*
[2]*Mexico*

1. Introduction

It is well known in the wireless telecommunications field that the most valuable resource available is the electromagnetic radio spectrum. Being a natural resource, it is obviously finite and has to be utilized in a rational fashion. Nevertheless the demand increase on wireless devices and services such as voice, short messages, Web, high-speed multimedia, as well as high quality of service (QoS) applications has led to a saturation of the currently available spectrum. On the other hand, it has be found that some of the major licensed bands like the ones used for television broadcasting are severely underutilized Federal Communications Commission (November) which at the end of the day results in a significant spectrum wastage. For this means, it is important to come up with a new paradigm that allows us to take advantage of the unused spectrum. Cognitive radio has risen as a solution to overcome the spectrum underutilization problem Mitola & Maguire (1999),Haykin (2005). The main idea under cognitive radio systems is to allow unlicensed users or cognitive users (those who have not paid for utilizing the electromagnetic spectrum), under certain circumstances, to transmit within a licensed band. In order to perform this task, cognitive users need to continuously monitor the spectrum activities and find a suitable spectrum band that allows them to:

- Transmit without or with the minimum amount of interference to the licensed or primary users.
- Achieve some minimum QoS required for their specific application.
- Share the spectrum with other cognitive users.

Therefore, it is easy to observe that spectrum sensing is the very task upon which the entire operation of cognitive radio rests Haykin et al. (2009). It is of extreme importance for the system to be able to detect the so-called spectrum holes (underutilized subbands of the radio spectrum). This is why in this chapter we focus all our attention to analyze some important aspects of spectrum sensing in cognitive radio, and particularly the case when it is performed using multiple antennas.

In order to take advantage of the cognitive radio features it is important to find which parts of the electromagnetic spectrum are unused at certain moment. These portions are also called *spectrum holes* or *white spaces*. If these bands are further used by a licensed user the cognitive radio device has the alternative of either moving to another spectrum hoe or staying in the same band but altering its transmission power lever or modulation scheme in order to avoid the interference. Hence it is clear that an important requirement of any cognitive radio network is the ability to sense such spectrum holes. As the most recent literature suggests right now Akyildiz et al. (2008),Haykin et al. (2009), the most efficient way to detect spectrum holes is to detect the primary users that are receiving data within the communication range of a cognitive radio user. This approach is called *transmitter detection* which is based on the detection of the weak signal from a primary transmitter through the local observations of cognitive users. The hypotheses cab be defined as

$$x(t) = \begin{cases} \mathcal{H}_0 : & n(t) \\ \mathcal{H}_1 : hs(t) + n(t) \end{cases}, \tag{1}$$

where $x(t)$ is the signal received by the cognitive user, $s(t)$ is the transmitted signal of the primary user, $n(t)$ is the AWGN and h is the amplitude gain of the channel. $\mathcal{H}_0$ is a null hypothesis, which states that there is no licensed user signal in a certain spectrum band. On the other hand, $\mathcal{H}_1$ is an alternative hypothesis, which states that there exist some licensed user signal. Three very famous models exist in order to implement transmitter detection according to the hypotheses model Poor & Hadjiliadis (2008). These are the matched filter detection, the energy detection and the cyclostationary feature detection.

1.1 Matched filter detection

When the information about the primary user signal is known to the cognitive user, the optimal detector in stationary Gaussian noise is the matched filter since it maximizes the received signal to noise ratio (SNR). While the main advantage of the matched filter is that it requires less time to achieve high processing gain due to coherency, it requires a priori knowledge of the primary user signal such as the modulation type and order, the pulse shape and the packet format. So that, if this information is not accurate, then the matched filter performs poorly. However, since most wireless networks systems have pilot, preambles, synchronization word or spreading codes, these can be used for coherent detection,

1.2 Energy detection

If the receiver cannot gather sufficient information about the primary user signal, for example, if the power of the random Gaussian noise is only known to the receiver, the optimal detector is an energy detector. In order to measure the energy of the received signal, the output signal of bandpass filter with bandwidth W is squared and integrated over the observation interval T. Finally, the output of the integrator Y, is compared with some threshold λ to decide whether a licensed user is present or not. Nevertheless, the performance of the energy detector is very susceptible to uncertainty in noise power. Hence, in order to solve this problem, a pilot tone from the primary transmitter can be used to help improve the accuracy of the energy detector. Another shortcoming is that the energy detector cannot differentiate signal types but can only determine the presence of the signal. Thus the energy detectors is prone to the false detection triggered by the unintended signals.

1.3 Cyclostationary feature detection

An alternative detection method is the cyclostationary feature detection. Modulated signals are in general couple with sine wave carriers, pulse trains, repeating spreading, hopping sequences or cyclic prefixes, which result in built-in periodicity Kontorovich et al. (2010). These modulated signals are characterized as cyclostationary since their mean and autocorrelation exhibit periodicity. These features are detected by analyzing a spectral correlation function. The main advantage of the spectral correlation function is that it differentiates the noise energy from modulated signal energy, which is a result of the fact that the noise is a wide-sense stationary signal with no correlation, while modulated signals are cyclostationary with spectral correlation due to the embedded redundancy of signal periodicity. Therefore, a cyclostationary feature detector can perform better than the energy detector in discriminating against noise due to its robustness to the uncertainty in noise power. Nonetheless, it is computationally complex and requires significantly long observation time.

Most of the previously mentioned techniques are investigated for a single sensor albeit some use of multiple sensors is suggested in (Zhang et al., 2010). In the latter, the authors consider a single sensor scenario equipped with multiple antennas and derived its performance in assumption of correlated antennas and constant channel. Also, most of these studies are focused on investigating the performance of particular schemes in ideal environments such as independent antennas in cooperative scenario or in uniform scattering. However, such consideration eliminate impact of real environment and its variation, while it is shown in many publications and realistic measurements that such environments change frequently, especially in highly build areas. Understanding how particular radio environment affects performance of cognitive radio sensing abilities is, therefore, and important issue to consider. Furthermore, it is well known (Haghighi et al., 2010) that the distribution of angle of arrival (AoA), itself defined by scattering environment (Haghighi et al., 2010), affects both temporal and spatial correlation of signals in antenna arrays. For these reasons in the first part of the chapter we utilize a simple but generic model of AoA distribution, suggested in (Abdi & Kaveh, 2002), to describe impact of scattering on statistical properties of received signals. Later the concept of Stochastic Degrees of Freedom (SDoF) is incorporated in order to obtain approximate expressions for the probability of miss detection in terms of number of antennas, scattering parameters and number of observations. Following, the trade-off between the number of antennas and required observation duration in correlated fading environments is investigated. It is shown that at low SNR it is more convenient having just a single antenna and many time samples so the noise suppression performs better. On the contrary, at high SNR, since the noise is suppressed relatively quickly is better to have more antennas in order to mitigate fading. Now, most of the existing spectrum schemes are based on fixed sample size detectors, which means that their sensing time is preset and fixed. Hence, in the second part of the chapter, we present some results based on the work of A. Wald (Wald, 2004) which showed that a detector based on sequential detection requires less average sensing time than a fixed size detector. We show that in general, it is possible to achieve the same performance that other fixed sample based techniques offer but using as low as half of the samples in average in the low signal to noise ratio regime. Afterwards, the impact of non-coherent detection is assessed when detecting signals using sequential analysis. We finished using sequential analysis as a new approach of cooperative approach for sensing. We call this an optimal fusion rule for distributed Wald detectors and a evaluate its performance. The last section of the chapter is devoted to conclusion remarks.

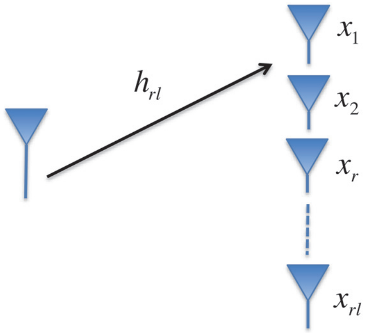


Fig. 1. System Model

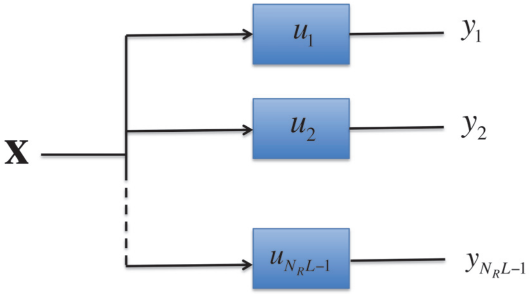


Fig. 2. Filtered Observations

2. Impact of scattering environment in spectrum sensing in multi-antenna cognitive radio systems

2.1 Signal model

Let us consider a primary transmitter which transmits some pilot signal s over L symbols in order to sound the primary channel. CR can sense the same signal using N_R receiving antennas. The received signal matrix $\mathbf{X}$ of size $N_R \times L$ can be written in terms of the $N_R \times L$ complex channel matrix $\mathbf{H} = \{h_{rl}\}$ and the noise matrix $\mathbf{W}$ of the same size as

$$\mathbf{X} = \mathbf{H}s + \mathbf{W}, \tag{2}$$

Here $\mathbf{W}$ is a zero mean Gaussian matrix of covariance $\sigma_n \mathbf{I}$ and $\mathbf{H}$ is a zero mean Gaussian matrix with covariance matrix $\mathbf{R_H}$ respectively. Element h_{rl} is the channel transfer coefficient from the transmitter to r-th antenna measured at l-th pilot symbol. Using vectorization operation ((van Trees, 2001)) , one can rewrite (2) as

$$\mathbf{x} = \mathbf{h}s + \mathbf{w}, \tag{3}$$

where $\mathbf{x} = \operatorname{vec}\mathbf{X}$, $\mathbf{h} = \operatorname{vec}\mathbf{H}$ and $\mathbf{w} = \operatorname{vec}\mathbf{W}$[1]. Therefore, the detection problem is to distinguish between the hypotheses

$$\begin{array}{ll} \mathcal{H}_0 : x[n] = \quad w[n] & n = 0, 1, \ldots, N_R L - 1 \\ \mathcal{H}_1 : x[n] = h[n]s + w[n] & n = 0, 1, \ldots, N_R L - 1 \end{array}. \tag{4}$$

[1] The vec$(\cdot)$ operator is defined as the $N_R L \times 1$ vector formed by stacking the columns of the $N_R \times L$ matrix i.e. $\operatorname{vec}\mathbf{H} = [\mathbf{h}_1' \mathbf{h}_2' \ldots \mathbf{h}_L']'$

The sufficient statistic in this case is given by (van Trees, 2001), (Kay, 1998)

$$\mathcal{T} = \mathbf{x}^H \mathbf{Q} \mathbf{x} = |s|^2 \mathbf{x}^H \mathbf{R_h} \left[|s|^2 \mathbf{R_h} + \sigma_n^2 \mathbf{I} \right]^{-1} \mathbf{x}, \tag{5}$$

where $\mathbf{R_h} = \mathcal{E}\left\{\mathbf{h}\mathbf{h}^H\right\}$ is the correlation matrix of the channel vector $\mathbf{h} = \operatorname{vec}\mathbf{H}$. This correlation matrix reflects both spatial correlation between different antennas and the time-varying nature of the channel. Let $\mathbf{R_h} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H$ be eigendecomposition of the correlation matrix $\mathbf{R_h}$. In this case, the statistic $\mathbf{T}$ could be recast in terms of the elements of the eigenvalues λ_i of the matrix $\boldsymbol{\Lambda}$ and filtered observations $\mathbf{y} = \mathbf{U^H}\mathbf{x}$:

$$\mathcal{T} = \mathbf{y}^H \boldsymbol{\Lambda} \left[\boldsymbol{\Lambda} + \sigma_n^2 \mathbf{I} \right]^{-1} \mathbf{y} = \sum_{k=1}^{N_R L} \frac{\lambda_k^2}{\lambda_k^2 + \sigma_n^2} |y_k|^2, \tag{6}$$

which is analogous to equation (5.9) in (van Trees, 2001). Elements y_k of the vector $\mathbf{y}$ could be considered as filtered version of the received signal $\mathbf{x}$ with a set of orthogonal filters $\mathbf{u}_k$ (columns of the matrix $\mathbf{U}$), *i.e.* could be considered as multitaper analysis (Thomson, 1982). Linear filtering preserve Gaussian nature of the received signals, therefore, the distribution of $\mathcal{T}$ could be described by generalized χ^2 distribution[2] (Andronov & Fink, 1971):

$$p(x) = \sum_{k=1}^{N_R L} \alpha_k \exp(-x/2\lambda_k), \tag{7}$$

and

$$\alpha_k^{-1} = 2\lambda_k \prod_{l=1, l\neq k}^{N_R L} \left(1 - \frac{\lambda_l}{\lambda_k}\right). \tag{8}$$

Theoretically, equation (7) could be used to set up the detection threshold γ. However, it is difficult to use it for analytical investigation. Therefore, we would consider a few particular cases of the channel when the structure of the correlation matrix could be greatly simplified to reveal its effect on the detection performance.

2.2 Performance of estimator-correlator for PU detection

2.2.1 Constant independent channels

In this case the full covariance matrix $\mathbf{R_h} = \sigma_h^2 \mathbf{O}_L \otimes \mathbf{I}_{N_R}$ is a Kronecker product of $N_R \times N_R$ identity correlation matrix $\mathbf{I}_{N_R}$ and $\mathbf{O}_L = \mathbf{1}\mathbf{1}^H$ is a $L \times L$ matrix consisting of ones. Therefore, there are N_R eigenvalues λ_k, $k = 1, \cdots N_R$ equal to L. The k-th orthogonal filter $\mathbf{u}_k$ is the averaging operator applied to the data collected from the k-th antenna. Thus, the decision statistic is just

$$\mathcal{T}_{CI} = \sum_{k=1}^{N_R} \left| \sum_{l=1}^{L} x_{kl} \right|^2 = \sum_{k=1}^{N_R} P_k, \tag{9}$$

where

$$P_k = \left| \sum_{l=1}^{L} x_{il} \right|^2. \tag{10}$$

[2] Assuming that all eigenvalues λ_k of $\mathbf{R_h}$ are different.

In absence of the signal, samples x_{kl} are drawn from an i.i.d. complex Gaussian random variable with zero mean and variance σ_n^2. Therefore, the distribution of P_k is exponential, with the mean value $L\sigma_n^2$

$$p(P) = \frac{1}{L\sigma_n^2} \exp\left(-\frac{P}{L\sigma_n^2}\right), \tag{11}$$

and the distribution of $\mathcal{T}$ is just central χ^2 distribution with N_R degrees of freedom

$$p_{CI}(\mathcal{T}|\mathcal{H}_0) = \frac{1}{\Gamma(N_R)} \frac{\mathcal{T}^{N_R-1}}{(L\sigma_n^2)^{N_R}} \exp\left(-\frac{\mathcal{T}}{L\sigma_n^2}\right). \tag{12}$$

If γ is a detection threshold for the statistic $\mathcal{T}$ then the probability P_{FA} of the false alarm is

$$P_{FA} = \int_{\gamma}^{\infty} p_{CI}(\mathcal{T}|\mathcal{H}_0) d\mathcal{T} = \frac{\Gamma\left[N_R, \gamma/L\sigma_n^2\right]}{\Gamma(N_R)}, \tag{13}$$

or

$$\gamma_{CI} = L\sigma_n^2 \Gamma^{-1}\left[N_R, P_{FA}\Gamma(N_R)\right], \tag{14}$$

where $\Gamma^{-1}\left[N_R, \Gamma(N_R, x)\right] = x$. If the signal is present, *i.e.* if the hypothesis $\mathcal{H}_1$ is valid, then the signal y_i is a zero mean with the variance $\sigma^2 = L^2|s|^2\sigma_h^2 + L\sigma_n^2$. As the result, the distribution of the test statistic $\mathcal{T}$ under the hypothesis $\mathcal{H}_1$ is given by the central χ^2 distribution with N_R degree of freedom and the probability of the detection is just

$$P_D = \int_{\gamma}^{\infty} p(\mathcal{T}|\mathcal{H}_1) = \frac{1}{\Gamma(N_R)} \Gamma\left(N_R, \frac{\gamma}{\sigma^2}\right) = \frac{1}{\Gamma(N_R)} \Gamma\left(N_R, \frac{1}{1+L\bar{\mu}} \Gamma^{-1}\left[N_R, P_{FA}\Gamma(N_R)\right]\right), \tag{15}$$

where

$$\bar{\mu} = |s|^2 \frac{\sigma_h^2}{\sigma_n^2}, \tag{16}$$

is the average SNR per symbol. Performance curves for this case could be found in (van Trees, 2001).

It could be seen from both (9) and (15) that under the stated channel model, the improvement in performance P_D comes either through reduction of noise through accumulation of signal in each of the antennas (*i.e.* increase in the effective SNR) or through exploitation of diversity provided by N_R antennas Kang et al. (2010). Thus, increasing number of antennas leads to a faster detection.

2.2.2 Spatially correlated block fading (constant spatially correlated channel)

Now let us assume that the values of the channel remain constant over L symbols but the values of the channel coefficients for different antennas are correlated. In other words we will assume that $\mathbf{R_h} = \sigma_h^2\, \mathbf{O}_L \otimes \mathbf{R}_s$ where $\mathbf{R}_s$ is the spatial correlation matrix between antennas. Let $\mathbf{R}_s = \mathbf{U}_s \mathbf{\Lambda}_s \mathbf{U}_s^H$ be spectral decomposition of $\mathbf{R}_s$. Then the test statistic $\mathcal{T}$ could be expressed, according to equation (6), as

$$\mathcal{T}_{CC} = \sum_{k=1}^{N_R} \frac{|s|^2\sigma_h^2\lambda_k}{|s|^2\sigma_h^2\lambda_k + \sigma_n^2} |y_k|^2 = \sum_{k=1}^{N_R} \frac{\bar{\mu}\lambda_k}{\bar{\mu}\lambda_k + 1} |y_k|^2, \tag{17}$$

where σ_h^2 is the variance of the channel per antenna. The eigenvalues λ_k of $\mathbf{R}_s$ reflect time accumulation of SNR in each "virtual branch" of the equivalent filtered value y_k. In general, all the eigenvalues are different so one should utilize equation (7). While these calculations are relatively easy to implement numerically, it gives little insight into the effect of correlation on the performance of the detector.

Under certain scattering conditions (Haghighi et al., 2010), the eigenvalues of the matrix $\mathbf{R}_s$ are either all close to some constant $\lambda > 1$ or close to zero. If there is $N_{eq} < N_R$ non-zero eigenvalues, their values are equal to $\lambda_k = N_R/N_{eq}$ to preserve trace, and the rest $N_R - N_{eq}$ are equal to zero. In this case, the test statistic $\mathcal{T}_{CC}$ could be further simplified to

$$\mathcal{T}_{CC}(N_{eq}) = \sum_{k=1}^{N_{eq}} |y_k|^2, \tag{18}$$

where the index k corresponds to non-zero eigenvalues. Thus, the problem is equivalent to one considered in Section 2.2.1 with N_{eq} independent antennas and the expression for the threshold γ_{CC} and the probability of detection are given by

$$\alpha\gamma_{CC} = \sigma_n^2 \Gamma^{-1}\left[N_{eq}, P_{FA}\Gamma(N_{eq})\right], \tag{19}$$

where $0 < \alpha < 1$ performs as a corrector variable.

The effect of correlation between branches has dual effect on performance of the system. On one side, the number N_{eq} of equivalent independent branches is reduced, comparing to the number of antennas N_R, therefore reducing diversity. However, increased correlation results into additional accumulation of SNR (or, equivalently, additional noise reduction through averaging) by factor of $N_R/N_{eq} \geq 1$. Therefore

$$\begin{aligned} P_D = \int_{\gamma}^{\infty} p(\mathcal{T}|\mathcal{H}_1)d\mathcal{T} = \frac{1}{\Gamma(N_{eq})}\Gamma\left(N_{eq}, \frac{\alpha L \gamma_{CC}}{\sigma^2}\right) = \\ \frac{1}{\Gamma(N_{eq})}\Gamma\left(N_{eq}, \frac{1}{1 + LN_R\bar{\mu}/N_{eq}}\Gamma^{-1}\left[N_{eq}, P_{FA}\Gamma(N_{eq})\right]\right). \end{aligned} \tag{20}$$

2.2.3 Independent channel with temporal correlation

In the case of independent antennas but temporally correlated fading, the full correlation matrix can be represented as $\mathbf{R}_h = \mathbf{R}_T \otimes \mathbf{I}_L$ where $\mathbf{R}_T = \mathbf{U}_T^H \mathbf{\Lambda}_T \mathbf{U}$ is the eigen decomposition temporal correlation matrix of an individual channel Paulraj et al. (2003). The decision statistic can now be represented as

$$\mathcal{T}_{ICC} = \sum_{k=1}^{N_R} \mathbf{x}_k^H \mathbf{R}_T \left(\mathbf{R}_T + \frac{1}{\bar{\mu}}\mathbf{I}_L\right)^{-1} \mathbf{x}_k = \sum_{k=1}^{N_R} \mathcal{T}_k, \tag{21}$$

where $\mathbf{x}_k$ is $1 \times L$ time sample received by the k-th antenna. Therefore, each antenna signal is processes separately and the results are added afterwards.

Taking advantage of eigendecomposition of the correlation matrix $\mathbf{R}_T$ calculation of decision statistic $\mathcal{T}_k$ can be recast as a multitaper analysis

$$\mathcal{T}_k = \mathbf{y}_k \mathbf{\Lambda}_k \left(\mathbf{\Lambda}_k + \frac{1}{\bar{\mu}}\mathbf{I}_L\right)^{-1} \mathbf{y}_k = \sum_{l=1}^{L} \frac{\lambda_l}{\lambda_l + 1/\bar{\mu}} |y_{kl}|^2. \tag{22}$$

Once again, we can utilize approximation of the correlation matrix by one with constant or zero eigenvalues as in Section 2.2.2. In this case there will be

$$L_{eq} = \frac{(\mathrm{tr}\,\mathbf{R}_T)^2}{\mathrm{tr}\,\mathbf{R}_T\mathbf{R}_T^H} \tag{23}$$

eigenvalues of size L/L_{eq} and the rest are zeros. Therefore, there is $N_R L_e q$ terms in the sum (21) each contributing

$$\frac{L/L_{eq}}{L/L_{eq} + 1/\bar{\mu}} = \frac{\bar{\mu}L + L_{eq}}{\bar{\mu}L}, \tag{24}$$

into the variance of $\mathcal{T}_{ICC}$. Corresponding equations for choosing the threshold become

$$\gamma_{CC} = L\sigma_n^2 \Gamma^{-1}\left[N_R L_{eq}, P_{FA}\Gamma(N_R L_{eq})\right], \tag{25}$$

$$P_D = \int_{\gamma}^{\infty} p(\mathcal{T}|\mathcal{H}_1)d\mathcal{T} = \frac{1}{\Gamma(N_{eq})}\Gamma\left(N_{eq}, \frac{\gamma}{\sigma^2}\right) = \frac{1}{\Gamma(N_{eq})}\Gamma\left(N_{eq}, \frac{1}{1 + LN_R\bar{\mu}/N_{eq}}\Gamma^{-1}\left[N_{eq}, P_{FA}\Gamma(N_{eq})\right]\right). \tag{26}$$

2.2.4 Channel with separable spatial and temporal correlation

The correlation matrix of the channel with separable temporal and spatial correlation has the correlation matrix of the form $\mathbf{R_h} = \mathbf{R}_T \otimes \mathbf{R}_s$. Correlation in both coordinates reduces total number of degrees of freedom from $N_R L$ to $N_{eq}L_{eq} \leq N_R L$ The loss of degrees of freedom is offset by accumulation of SNR due to averaging over the correlated samples. The equivalent increase in the average SNR is $N_R L/N_{eq}L_{eq}$. Thus, the problem is equivalent to detection using

$$K_{eq} = N_{eq}L_{eq} = \frac{(\mathrm{tr}\mathbf{R}_s)^2}{||\mathbf{R}_s||_F^2}\frac{(\mathrm{tr}\mathbf{R}_T)^2}{||\mathbf{R}_T||_F^2}, \tag{27}$$

independent samples in the noise with the average SNR

$$\bar{\mu}_{eq} = \frac{N_R L}{N_{eq}L_{eq}}\bar{\mu}. \tag{28}$$

The sufficient statistics in the case of the channel with separable spatial and temporal correlation could be easily obtained from the general expression (5) and (6). In fact, using Kronecker structure of $\mathbf{R_h}$ one obtains

$$\mathcal{T} = \sum_{k=1}^{K_{eq}} |z_k|^2. \tag{29}$$

2.3 Examples and simulation

2.3.1 Correlation models

While the Jakes correlation function $J_0(2\pi f_D \tau)$ is almost universally used in standards on wireless channels (Editors, 2006), realistic environment is much more complicated. A few other models could be found in the literature, some chosen for their simplicity, some are based on the measurements. In most cases we are able to calculate N_{eq} analytically, as shown below.

1. **Sinc type correlation** If scattering environment is formed by a single remote cluster (as it is shown in (Haghighi et al., 2010)), then the spatial covariance function $\mathbf{R}_s(d)$ as a function of electric distance between antennas d is given by

$$\mathbf{R}_s(d) = \exp\left(j2\pi d \, \sin\phi_0\right) \operatorname{sinc}\left(\Delta\phi d \, \cos\phi_0\right), \tag{30}$$

where ϕ_0 is the central angle of arrival, $\Delta\phi$ is the angular spread. This correlation matrix has approximately $\lfloor 2\Delta\phi \cos\phi_0 N + 1 \rfloor$ eigenvalues approximately equal eigenvalues with the rest close to zero (Slepian, 1978).

2. **Nearest neighbour correlation** Neglecting correlation between any two antennas which are not neighbours one obtains the following form of the correlation matrix $\mathbf{R}_s$

$$\mathbf{R}_s = \left\{r_{ij}\right\} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i = j+1 \\ \rho^* & \text{if } i = j-1 \\ 0 & \text{if } |i-j| > 1 \end{cases}, \tag{31}$$

where ρ is the correlation coefficient. The eigenvalues of (31) are well know (Kotz & Adams, 1964)

$$\lambda_k = 1 - 2|\rho| \cos\frac{k\pi}{N+1}, \quad 1 \le k \le N. \tag{32}$$

The equivalent number of independent virtual antennas is given by

$$N_{eq} = \frac{N^2}{N + 2(N-1)|\rho|^2} = \frac{N}{1 + 2|\rho|^2\left(1 - 1/N\right)}. \tag{33}$$

3. **Exponential correlation**

$$\mathbf{R}_s = \left\{r_{ij}\right\} = \left\{|\rho|^{i-j}\right\}. \tag{34}$$

Eigenvalues of this matrix are well known (34) (Kotz & Adams, 1964)

$$\lambda_k = \frac{1 - |\rho|^2}{1 + 2|\rho| \cos\psi_k + |\rho|^2}, \tag{35}$$

where ψ_k are roots of the following equation

$$\frac{\sin(N+1)\psi - 2|\rho|\psi \sin N + |\rho|^2 \sin(N-1)\psi}{\sin\psi} = 0. \tag{36}$$

4. **Temporal correlation model for nonisotropic scattering** Considering the extended case of the Clarke's temporal correlation model for the case of nonisotropic scattering around the user, we have the temporal correlation function as (Abdi & Kaveh, 2002):

$$\mathbf{R}_s(\tau) = \frac{I_0\left(\sqrt{\kappa^2 - 4\pi^2 f_D^2 \tau^2 + j4\pi\kappa\cos(\theta) f_D \tau}\right)}{I_0(\kappa)}, \tag{37}$$

where $\kappa \ge 0$ controls the width of angle of arrival (AoA), f_d is the Doppler shift, and $\theta \in [-\pi, \pi)$ is the mean direction of AoA seen by the user; $I_0(\cdot)$ stands for the zeroth-order modified Bessel function.

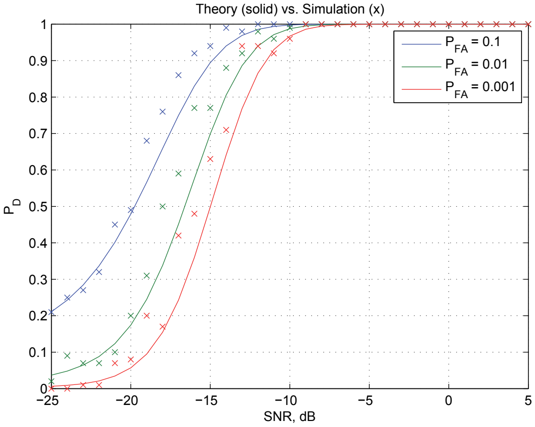


Fig. 3. ROC approximation vs. simulation results ($\alpha = 0.8$) . Solid lines - theory, x-lines - simulation.

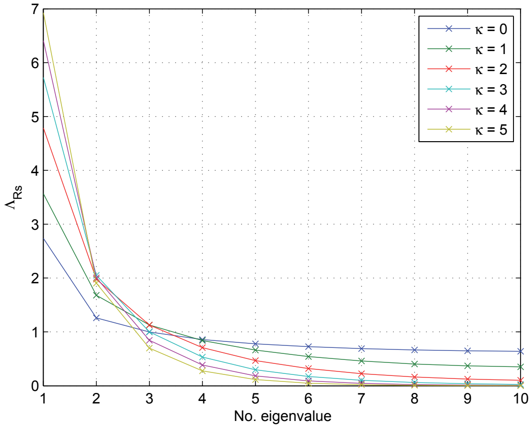


Fig. 4. Eigenvalues behavior of R_s temporal correlation matrix for nonisotropic scattering ($N = 10, \mu = 0$ and $f_d = 50Hz$)

Figure 4 shows the eigenvalues behaviour for different values of the κ factor. Notice that for $\kappa = 0$ (isotropic scattering) the values of the eigenvalues are spread in an almost equally and proportional fashion among all of them. As κ tends to infinity (extremely nonisotropic scattering), we obtain $N - 1$ zero eigenvalues and one eigenvalue with value N. In other words, as κ increases, the number of "significant" eigenvalues decreases and hence so the value of N_{eq} as shown in Figure 6.

2.4 Simulation procedure

In order to perform the simulations which verified these results, the hypothesis in eq. (4) was formed by giving the channel matrix $\mathbf{H}$ the desired correlation characteristics as shown in

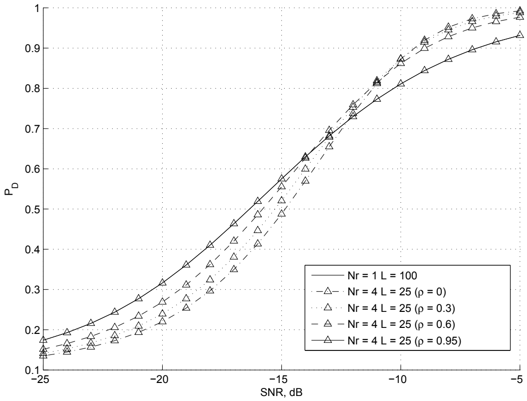


Fig. 5. Effect of correlation between antennas in the probability of detection.

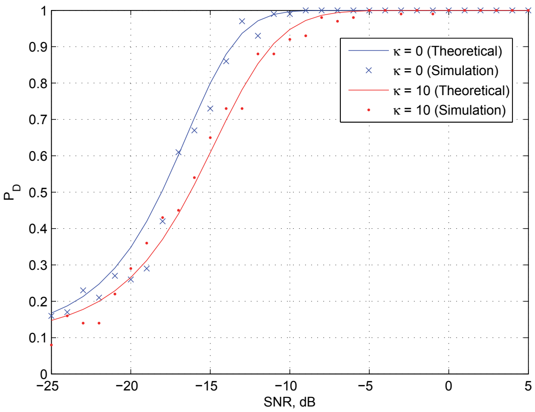


Fig. 6. ROC approximation for the estimator correlator considering the temporal correlation for isotropic and nonisotropic scattering ($\kappa = 0$ and $\kappa = 10$ respectively).

section 2.3.1. Therefore, the vectorization operations are performed and after evaluating the respective statistical tests, Monte Carlo method is utilized.

2.5 Space-time processing trade-off

It is common to assume that increasing number of antennas improves performance of detection algorithms due to increased degree of diversity. Such proposition is correct when the number of time samples remains the same. However, in cognitive networks it is desired to reduce decision time as much as possible, sometimes by introducing some added complexity in the form of additional number of antennas. The goal of this section is to show how one can trade speed of making decision with a number of antennas available for signal reception.

It can be seen from equation (18) that the processing of the signal consists of two separated procedures: averaging in time and accounting for diversity and suppressing noise in spatial diversity brunches. Depending on amount of noise (SNR) and fading (fading figure (Simon & Alouini, 2000)) one of these two technique brings more benefit to the net procedure. For relatively low levels of SNR noise suppression is a dominant task, therefore it is more advantageous to have a single antenna and as many time samples as possible. On a contrary, if SNR is somewhat higher, noise is sufficiently suppressed even by short time averaging and suppressing fading through diversity combining is more beneficial. This can be seen from Fig. 5. This graph shows performance different configurations of the receiver in such a way that product $N_R L$ remains constant. The same Figure shows effect correlation, and thus, the scattering environment, plays on quality of reception. For very strong correlations $\rho \approx 1$ and $N_{eq} \approx 1$. Therefore all samples collected are used to reduce noise. Such scheme performs the best at low SNR. However, when $\rho = 0$ and $N_{eq} = N_R$ the gain from the diversity is highest and the scheme performs best for higher SNR. Intermediate case allows for smooth transition between these two regions. It also can be seen, that only in the case of $\rho = 1$ there is an equivalent trade off between number of antennas and time samples: *i.e.* the performance depends only on $Q = N_R L$ regardless how the product is divided between N_R and L. However, lesser correlation result in unequal trade-off with gain or loss defined by SNR and amount of correlation.

3. Sequential analysis for multi-antenna cognitive radio

3.1 Sequential analysis overview

The sequential analysis and the sequential probability ratio test (SPRT) introduced by A. Wald in 1943 (Wald, 2004) have proved to be highly effective in taking decisions between two known hypothesis $(\mathcal{H}_0, \mathcal{H}_1)$. Moreover, as it is shown in (Wald, 2004), the sequential probability ratio test frequently results in a saving of about 50 percent in the average number of observations in comparison to other well known detection techniques such as the Neyman-Pearson (NP) decision test which is based on fixed number of observations. Unlike NP test, which utilizes the logarithm of the Likelihood Ratio (log-LR) and compares it with a single predefined threshold γ, the sequential probability ratio test compares the log-LR with two thresholds $\mathcal{A}$ and $\mathcal{B}$ which are obtained in terms of the target probability of false alarm (P_{FA}) and probability of misdetection (P_{MD}) (or the complementary probability of detection $P_D = 1 - P_{MD}$) (Wald, 2004), (Middleton, 1960). Furthermore, in contrast to fixed decision time of NP test, the duration of testing of sequential analysis is a random variable.

The thresholds $\mathcal{A}$ and $\mathcal{B}$ are approximated as (Wald, 2004):

$$\mathcal{A} = \ln \frac{1 - P_{MD}}{P_{FA}}, \ \mathcal{B} = \ln \frac{P_{MD}}{1 - P_{FA}} \tag{38}$$

The test procedure consists of sequential accumulating of m samples and calculating the cumulative sum of the m-th log-LR as

$$\mathcal{B} < \sum_{i=1}^{m} \Lambda_i < \mathcal{A}, \tag{39}$$

where Λ_i is a single log-likelihood ratio sample.

If Eq. (39) is satisfied, the experiment is continued by taking an additional sample increasing m by 1. However, if

$$\sum_{i=1}^{m} \Lambda_i \geq \mathcal{A}, \tag{40}$$

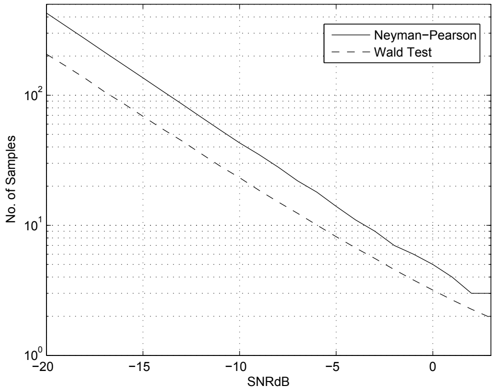


Fig. 7. Comparison of Neyman-Pearson Test and Sequential Probability Ratio Test ($P_{FA} = 0.1$,$P_D = 0.9$).

the process is terminated with the acceptance of $\mathcal{H}_1$. Similarly

$$\sum_{i=1}^{m} \Lambda_i \leq \mathcal{B}, \tag{41}$$

leads to termination with the acceptance of $\mathcal{H}_0$.

3.2 Wald test for complex random variable

Let us consider testing zero mean hypothesis in complex AWGN

$$\begin{array}{rl} \mathcal{H}_0 : z_i = & x_i + jy_i = w_i \\ \mathcal{H}_1 : z_i = & m + w_i \end{array}. \tag{42}$$

Here $m = m_I + jm_Q = \mu \exp(j\phi_m) \neq 0$ is complex non zero mean and w_i is i.i.d. complex zero mean Gaussian process of variance σ^2.
A single sample log-likelihood ratio Λ_i is given by

$$\Lambda_i = \ln \frac{p_1(z_i; \mathcal{H}_1)}{p_0(z_i; \mathcal{H}_0)} = \frac{2\mu\,(x_i \cos\phi_m + y_i \sin\phi_m) - \mu^2}{\sigma^2}. \tag{43}$$

After N steps of the sequential test the cumulative log-likelihood Λ ratio becomes

$$\Lambda = \sum_{n=1}^{N} \Lambda_n = \frac{2\mu}{\sigma^2}\mathcal{T}_N - \frac{N\mu^2}{\sigma^2}, \tag{44}$$

where

$$\mathcal{T}_N = \cos\phi_m \sum_{n=1}^{N} x_n + \sin\phi_m \sum_{n=1}^{N} y_n. \tag{45}$$

The rest of the test follows procedure of Section 3.1.

Figure 7 shows the performance comparison in number of samples needed between Wald Test and Neyman-Pearson Test. Notice that the latter needs in general almost twice the number of samples in order to detect the presence of the signal.
It follows from (45) that the sufficient statistic in the case of complex observation is

$$\mathcal{T} = \sum_{n=1}^{N} \Re\{x \exp(-j\phi_m)\}. \tag{46}$$

In other words, processing of the received signal is implemented in two stages: first, the data is unitary rotated by the angle ϕ_m to align the mean along real axis; then the real part of data is analyzed using the same procedure as purely real data.

3.2.1 Average number of samples

We have defined the sequential test procedure in eq.(39). Using this we can call

$$\Lambda = \sum_{i=1}^{m} \Lambda_i = \ln \frac{p(z_1, \ldots, z_m|\mathcal{H}_1)}{p(z_1, \ldots, z_m|\mathcal{H}_0)}, \tag{47}$$

where the random variable m stands for the required number of samples needed to terminate the test. As stated in Wald (2004) it is possible to neglect the excess on threshold $\mathcal{A}$ and $\mathcal{B}$, hence the the random variable can have the four possible combinations of terminations and hypotheses such as

$$\Lambda = \begin{cases} P_{FA}\mathcal{A} \text{ if } \mathcal{H}_0 \text{ is true} \\ P_D\mathcal{A} \text{ if } \mathcal{H}_1 \text{ is true} \\ (1 - P_{FA})\mathcal{B} \text{ if } \mathcal{H}_0 \text{ is true} \\ P_M\mathcal{B} \text{ if } \mathcal{H}_1 \text{ is true} \end{cases}. \tag{48}$$

Following same reasoning we can get the conditional expectation for the random variable Λ as

$$\bar{\Lambda} = \begin{cases} P_{FA}\mathcal{A} + (1 - P_{FA})\mathcal{B} & \text{if } \mathcal{H}_0 \text{ is true} \\ P_D\mathcal{A} + P_M\mathcal{B} & \text{if } \mathcal{H}_1 \text{ is true} \end{cases}. \tag{49}$$

It is possible now to obtain the average number of samples (decision time) for accepting one of the two hypothesis $(\mathcal{H}_0, \mathcal{H}_1)$ as:

$$\bar{n}(\mathcal{H}_0) = \frac{P_{FA}\mathcal{A} + (1 - P_{FA}\mathcal{B})}{\bar{\Lambda}(\mathcal{H}_0)}, \tag{50}$$

$$\bar{n}(\mathcal{H}_1) = \frac{P_D\mathcal{A} + (1 - P_D\mathcal{B})}{\bar{\Lambda}(\mathcal{H}_1)}, \tag{51}$$

where the term $\bar{\Lambda}(\mathcal{H}_0)$ can be calculated as

$$\bar{\Lambda}(\mathcal{H}_0) = \frac{\sum_{n=1}^{N} \Lambda_n}{N}, \tag{52}$$

if no signal is present. The term $\bar{\Lambda}(\mathcal{H}_1)$ can be calculated analogously assuming there is a signal present as follows

$$\bar{\Lambda}(\mathcal{H}_1) = \frac{\sum_{n=1}^{N} \Lambda_n}{N}. \tag{53}$$

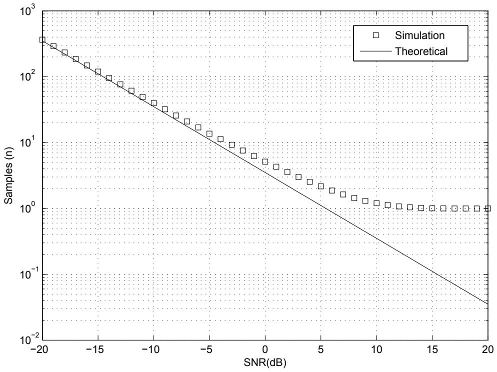


Fig. 8. Average Number of Samples for Detection in Sequential Analysis

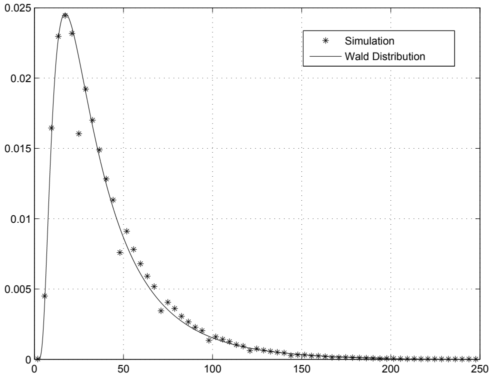


Fig. 9. Approximation of decision time using Wald Distribution.

Fig. 8 shows the average number of samples needed to achieve a $P_D = 0.9$ for different SNR. The deviation at high SNR's is due to the same effect already explained before. In practice for very high SNR's only one sample is enough to detect the presence of primary users.

3.2.2 Decision time distribution

as it was described earlier, decision time when using sequential analysis for detection is a random variable. Hence it can be described by its PDF. Although an exact shape for this PDF is not known in general, a very good approximation is available (specially for the low SNR regimen) called Wald distribution or inverse Gaussian distribution defined as

$$f(x) = \frac{\lambda}{2\pi x^3} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x} \qquad x > 0, \tag{54}$$

where μ is the mean and $\lambda > 0$ is the shape parameter. In figure 9 it is shown Wald's distribution in order to approximate the decision time for a $P_D = 0.9$.

Though application of sequential analysis with high reliability of the hypothesis testing (P_{FA}, $P_M \to 0$) can provide an effective censoring of the information sent to other SU or FC together with reduction of the sampling size at SU.

3.3 Sequential probability ratio test for partially coherent channel

Let us consider detection of a signal in a channel a with partially known phase. The received signal can be modeled as

$$z_i = m\exp(j\Delta) + w_i, \tag{55}$$

where $m = m_I + jm_Q = \mu\exp(j\phi_m)$ is a deterministic and known complex constant, w_i is complex zero mean Gaussian noise with variance σ^2. Random variable Δ represents uncertainty in measuring the phase of the carrier. Its distribution can be described by PDF $p_\Delta(\Delta)$. In the following we assume that the phase uncertainty is described by the Von Mises (or Tikhonov) PDF (von Mises, 1964)

$$p_\Delta(\delta) = \frac{\exp\left[\kappa\cos(\Delta - \Delta_0)\right]}{2\pi I_0(\kappa)}. \tag{56}$$

Parameter Δ_0 represents bias in the determination of the carrier's phase, while κ represents quality of measurements. A few particular cases could be obtained from (56) by proper choice of parameters

1. Perfect phase recovery (coherent detection): $\kappa = \infty$, $\Delta_0 = 0$, and, thus, $p_\Delta(\Delta) = \delta(\Delta)$
2. No phase recovery (non-coherent detection): $\kappa = 0$ and, $p_\Delta(\Delta) = 1/2\pi$
3. Constant bias: $\kappa = \infty$, $\Delta_0 \neq 0$, $p_\Delta(\Delta) = \delta(\Delta - \Delta_0)$

We will derive the general expression first and then investigate particular cases to isolate effects of the parameters on performance of SPRT.

3.3.1 Average likelihood ratio

For a single observation z_i probability densities $p_1(z_i)$ and $p_0(z_i)$ corresponding to each of the hypothesis $\mathcal{H}_1$ and $\mathcal{H}_0$ are given by

$$p_1(z_i) = C\exp\left[-\frac{(x_i - \mu\cos(\phi_m + \Delta))^2}{\sigma^2}\right]\exp\left[-\frac{(y_i - \mu\sin(\phi_m + \Delta))^2}{\sigma^2}\right], \tag{57}$$

and

$$p_0(z_i) = C\exp\left[-\frac{x_i^2 + y_i^2}{\sigma^2}\right], \tag{58}$$

respectively Filio, Primak & Kontorovich (2011). For a given Δ, the likelihood ratio L_i could be easily calculated to be

$$L_i = \frac{p_1(z_i)}{p_0(z_i)} = \exp\left[\frac{2\mu\left(x_i\cos(\phi_m + \Delta) + y_i\sin(\phi_m + \Delta)\right) - \mu^2}{\sigma^2}\right]. \tag{59}$$

The conditional (on Δ) likelihood ratio $L(N|\Delta)$, considered over N observation is then just a product of likelihoods of individual observations, therefore

$$L(N|\Delta) = \prod_{n=1}^{N} \frac{p_1(z_n)}{p_0(z_n)} = \exp\left[\frac{2\mu \sum_{n=1}^{N} (x_n \cos(\phi_m + \Delta) + y_n \sin(\phi_m + \Delta)) - N\mu^2}{\sigma^2}\right]$$
$$= \exp\left[\frac{2\mu \mathcal{T}(N,\Delta)}{\sigma^2}\right] \exp\left[-\frac{N\mu^2}{\sigma^2}\right], \quad (60)$$

where

$$\mathcal{T}(N,\Delta) = \cos(\phi_m + \Delta) \sum_{n=1}^{N} x_n + \sin(\phi_m + \Delta) \sum_{n=1}^{N} y_n. \quad (61)$$

Let us introduce a new variables, $X(N)$, $Y(N)$, $Z(N)$ and $\Psi(N)$, defined by

$$X(N) = Z(N) \cos \Psi(N) = \sum_{n=1}^{N} x_n \quad (62)$$

$$Y(N) = Z(N) \sin \Psi(N) = \sum_{n=1}^{N} y_n \quad (63)$$

Using this notation equation (61) can now be rewritten as

$$\mathcal{T}(N,\Delta) = Z(N) \cos\left[\phi_m + \Delta - \Psi(N)\right]. \quad (64)$$

The average likelihood (Middleton, 1960) $L(N)$ could now be obtained by averaging (60) over the distribution of $p_\Delta(\Delta)$ to produce

$$\bar{L}(N) = \exp\left[-\frac{N\mu^2}{\sigma^2}\right] \int_{-\pi}^{\pi} \exp\left[\frac{2\mu Z(N) \cos\left[\phi_m + \Delta - \Psi(N)\right]}{\sigma^2}\right] p_\Delta(\Delta) d\Delta. \quad (65)$$

In turn, this expression could be further specialized if $p_\Delta(\Delta)$ is given by equation (56)

$$\bar{L}(N) = \exp\left[-\frac{N\mu^2}{\sigma^2}\right] \frac{1}{I_0(\kappa)} I_0 \left[\sqrt{\frac{4\mu^2 Z^2(N)}{\sigma^4} + \frac{4\mu Z(N)\kappa}{\sigma^2} \cos\left[\phi_m - \Psi(N) - \Delta_0\right] + \kappa^2}\right]. \quad (66)$$

It can be seen from (66) that it is reduced to (45) if $\Delta_0 = 0$ and $\kappa = \infty$. Furthermore the deterministic phase bias Δ_0 could be easily eliminated from consideration by considering $\tilde{z}_i = z_i \exp[-j(\phi_m + \Delta_0)]$ instead of z_i. Therefore, equation (66) could be simplified to

$$\bar{L}(N) = \exp\left[-\frac{N\mu^2}{\sigma^2}\right] \frac{1}{I_0(\kappa)} I_0 \left[\frac{1}{\sigma^2} \sqrt{4\mu^2 Y^2(N) + [2\mu X(N) + \kappa\sigma^2]^2}\right]. \quad (67)$$

In the case of non-coherent detection, *i.e.* in the case $\kappa = 0$, expression (67) assumes a well known form

$$\bar{L}(N) = \exp\left[-\frac{N\mu^2}{\sigma^2}\right] I_0 \left[\frac{2\mu Z(N)}{\sigma^2}\right]. \quad (68)$$

Formation of the likelihood ratio could be considered as a two step process. As the first step, inphase and quadrature components independently accumulated to lessen the effect of AWGN. At the second step, values of $X(N)$ and $Y(N)$ must be combined in a fashion depending on available information. In the case of coherent reception it is know a priori that

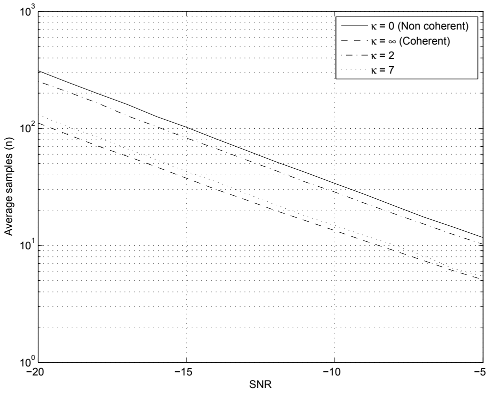


Fig. 10. Impact of coherency on the average number of samples.

the quadrature component $Y(N)$ contains only noise and it is ignored in the likelihood ratio. On the contrary, in the case of non-coherent reception one cannot distinguish between the in-phase and quadrature components and their powers are equally combined to for $Z(N)$. In the intermediate case both components are combined according to (67) with more and more emphasis put on in-phase component $X(N)$ as coherency increases with increase of κ.

In figure 10 we present the impact of non-coherent detection in the number of samples needed in order to detect a signal with respect to some P_D target. Notice that the main repercussion of non-coherence detection is the increase of samples to nearly twice in comparison to the coherent detector. In this terms, the non-coherent Wald sequential test procedure can be thought as having the same efficiency (in terms of number of samples) as the coherent NP-test.

3.4 Optimal fusion rule for distributed Wald detectors

This Section generalized results of Chair & Varshney (1986) to the case of distributed detection using Wal sequential analysis test. We assume that there are M sensors, making individual detection according to the Wald algorithm. Once a decision is made at an individual sensor the decision is send in the binary form to the Fusion Centre for further combining with other decisions. We assume that the value $u = -1$ is assigned if the hypothesis $\mathcal{H}_0$ is accepted, $u = 1$ if the hypothesis $\mathcal{H}_1$ is accepted and $u = 0$ if no decision has been made yet. Only $u = \pm 1$ are communicated to the Fusion Centre.

Since each node utilizes the Wald sequential detection Wald (2004), the decision is made at a random moments of time. Therefore, at any given moment of time t there is a random number $L(t) \leq M$ of decisions which are available at FC as can be seen in figure 11. Probability distribution of making decision could be approximated either by the two parametric Wald distribution Wald (2004), or by three parametric generalized inverse Gaussian distribution Jørgensen (1982) (as seen in fig.9). Parameters of those distributions could be found through moment/cumulant fitting, using expressions derived in Wald (2004), Filio, Kontorovich & Primak (2011). Following Chair & Varshney (1986) we would treat this problem as a two-hypothesis detection problem with individual detector decision being the observation. For a given number $L = L(t)$ of decisions made by the time t, the optimum decision rule is equivalent to the following likelihood ratio test

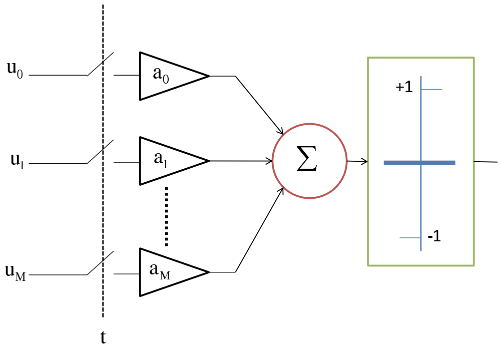


Fig. 11. System model of data fusion system

$$\frac{P\left(u_1,u_2,\cdots,u_L|L,\mathcal{H}_1\right)P(L|\mathcal{H}_1)}{P\left(u_1,u_2,\cdots,u_L|L,\mathcal{H}_0\right)P(L|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}. \quad (69)$$

Here $P(L|\mathcal{H}_0)$ is the probability of making exactly L decisions assuming that $\mathcal{H}_1$ is true and u_l is the decision made by l-th sensor.
Furthermore, assuming the minimum probability of error criteria, *i.e.* by setting $C_{00}=C_{11}=0$ and $C_{01}=C_{10}=1$, introducing notation $\mathbf{u}_L=\{u_1,u_2,\cdots,u_L\}$ and using the Bayes rule one can recast equation (69) as

$$\frac{P(\mathcal{H}_1|\mathbf{u}_L,\mathbf{l})P(\mathbf{l}|\mathcal{H}_1)}{P(\mathcal{H}_0|\mathbf{u}_L,\mathbf{l})P(\mathbf{l}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} 1, \quad (70)$$

or, after taking natural log of both side

$$\ln\frac{P(\mathcal{H}_1|\mathbf{u}_L,\mathbf{l})}{P(\mathcal{H}_0|\mathbf{u}_L,\mathbf{l})}+\ln\frac{P(\mathbf{l}|\mathcal{H}_1)}{P(\mathbf{l}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} 0, \quad (71)$$

where $\mathbf{l}$ is a vector which represents which sensors have made decisions.
Once again following Chair & Varshney (1986), one can calculate probabilities $P(\mathcal{H}_1|\mathbf{u}_L,\mathbf{l})$ and $P(\mathcal{H}_0|\mathbf{u}_L,\mathbf{l})$ as follows. In the case of the hypothesis $\mathcal{H}_1$ one can write

$$P(\mathcal{H}_1|\mathbf{u}_L,\mathbf{l})=\frac{P(\mathcal{H}_1,\mathbf{u}_L|\mathbf{l})}{P(\mathbf{u}_L|\mathbf{l})}=$$
$$\frac{P_1}{P(\mathbf{u}_L|\mathbf{l})}\prod_{S_+}P(u_l=+1|\mathcal{H}_1)\prod_{S_-}P(u_l=-1|\mathcal{H}_1)=\frac{P_1}{P(\mathbf{u}_L|\mathbf{l})}\prod_{S_+}(1-P_{M,l})\prod_{S_-}P_{M,l}, \quad (72)$$

where S_+ is the set of all i such that $u_i=+1$ and S_- is the set of all i such that $u_i=-1$.
Similarly, in the case of the hypothesis $\mathcal{H}_0$ one obtains

$$P(\mathcal{H}_0|\mathbf{u}_L,\mathbf{l})=\frac{P(\mathcal{H}_0,\mathbf{u}_L|\mathbf{l})}{P(\mathbf{u}_L|\mathbf{l})}=\frac{P_0}{P(\mathbf{u}_L|\mathbf{l})}\prod_{S_-}(1-P_{F,l})\prod_{S_+}P_{F,l}. \quad (73)$$

Finally, using equations (72) and (73) one obtains the following expression for the conditional log-likelihood

$$\ln\frac{P(\mathcal{H}_1|\mathbf{u}_L,\mathbf{l})}{P(\mathcal{H}_0|\mathbf{u}_L,\mathbf{l})}=\ln\frac{P_1}{P_0}+\sum_{S_+}\ln\frac{1-P_{M,l}}{P_{F,l}}+\sum_{S_-}\ln\frac{P_{M,l}}{1-P_{F,l}}. \quad (74)$$

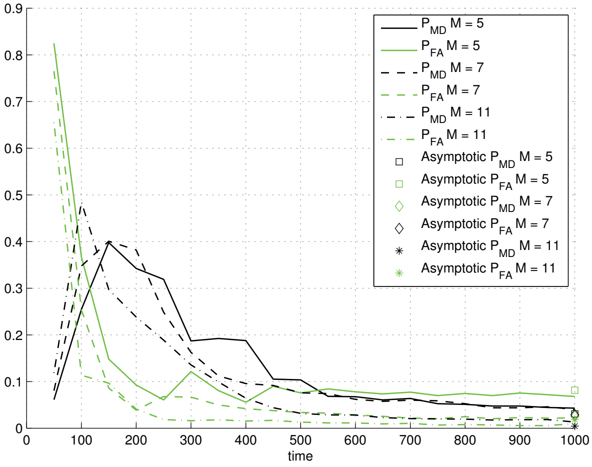


Fig. 12. Data fusion scheme considering sequential analysis decision from each sensor $(P_{M,l} = 0.3, P_{F,l} = 0.1)$.

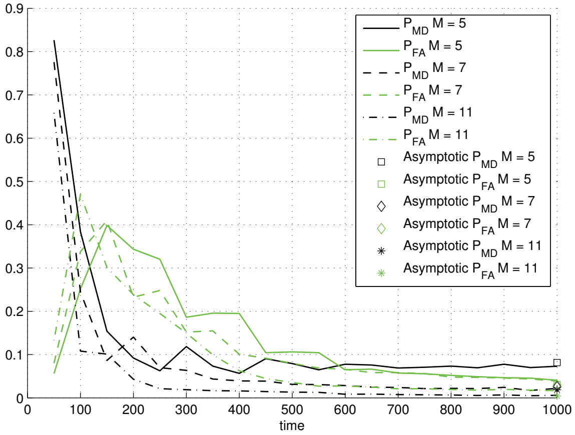


Fig. 13. Data fusion scheme considering sequential analysis decision from each sensor $(P_{M,l} = 0.1, P_{F,l} = 0.3)$.

In order to evaluate the second term in the sum (71) let us first consider an arbitrary node $1 \leq k \leq N$. Distribution $p_{T,k}(\tau)$ of the decision time at such a node is assumed to be known. Therefore, probability $P_{D,k}(t|\mathcal{H}_i)$ that the decision is made by the time $t = mT_s$ given that the hypothesis is true is given by

$$P_{D,k}(t|\mathcal{H}_i) = \int_0^{mT_S} p_k(\tau|\mathcal{H}_i)d\tau. \tag{75}$$

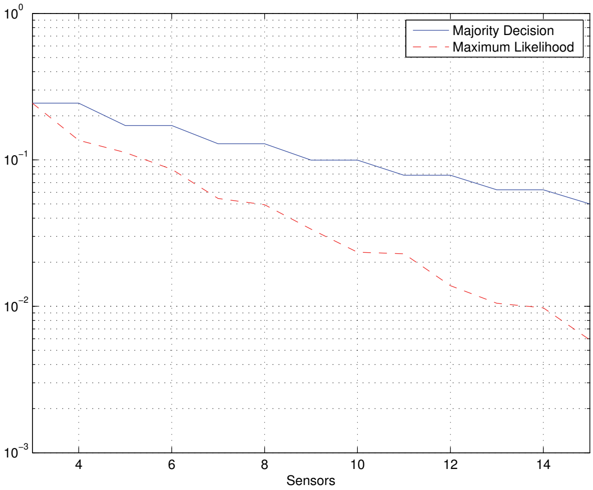


Fig. 14. Total Error Criteria

The probability that no decision has been made by the time t is then simply $1 - P_{D,k}(t|\mathcal{H}_i)$. As it has been mentioned earlier, parameters of this distribution could be defined as in Filio, Kontorovich & Primak (2011). Therefore , the second term in the equation (71).

$$\ln \frac{P(\mathbf{1}|\mathcal{H}_1)}{P(\mathbf{1}|\mathcal{H}_0)} = \sum_{l=1}^{L} \ln \frac{P_{D,l}(t|\mathcal{H}_1)}{P_{D,l}(t|\mathcal{H}_0)} + \sum_{l=L+1}^{M} \ln \frac{1 - P_{D,l}(t|\mathcal{H}_1)}{1 - P_{D,l}(t|\mathcal{H}_0)}. \tag{76}$$

Finally, the fusion rule in the case of nodes making decision according to Wald's criteria could be written as

$$f(\mathbf{u}) = \begin{cases} 1 & \text{if } a_0 + \sum_{l=1}^{L} a_l u_l > 0 \\ -1 & \text{otherwise} \end{cases} \tag{77}$$

where

$$a_0 = \ln \frac{P_1}{P_0} + \sum_{l=1}^{L} \ln \frac{P_{D,l}(t|\mathcal{H}_1)}{P_{D,l}(t|\mathcal{H}_0)} + \sum_{l=L+1}^{M} \ln \frac{1 - P_{D,l}(t|\mathcal{H}_1)}{1 - P_{D,l}(t|\mathcal{H}_0)}, \tag{78}$$

$$a_l = \ln \frac{1 - P_{M,l}}{P_{F,l}} \quad \text{if } u_l = 1, \tag{79}$$

$$a_l = \ln \frac{1 - P_{F,l}}{P_{M,l}} \quad \text{if } u_l = -1. \tag{80}$$

Thus, the combining rule is similar to that suggested in Chair & Varshney (1986), however, with some significant differences in the term of a_0. In figures 12 and 13 it is shown the performance of the data fusion scheme considering that each one of the sensors takes a decision based on the sequential detection criteria. In these figures it is plotted the probability of missdetection (P_{MD}) and the probability of false alarm (P_{FA}) versus the time in which the fusion centre gathers the decisions taken from the local observers. Notice that for $t \to \infty$ all graphs converge to the data fusion rule of Chair & Varshney (1986) for which the theoretical

expression for probability of missdetection and false alarm is derived in appendix 5. It is obvious that for small values of time, the fusion centre has less information (since not all the detectors might achieve a decision by then) and the final decision it takes is way less accurate that for large values of time. Nevertheless, in some practical systems it would be impossible to wait that long for getting the decision from the fusion centre, so we can use these results as a trade-off between the performance on the detection and the time of decision Gosan et al. (2010). Next thing it is possible to observe is the impact that $P_{D,l}$ and $P_{F,l}$ has on the performance of the data fusion detector.
Notice that for very small values of false alarm probability, $a_l \approx -\ln P_{M,l}$ if $u = -1$ in eq. (78) which means that hypothesis $\mathcal{H}_0$ is always less weighted in equation (77) or in other words, the fusion centre "trusts" more in those sensors who decide that $\mathcal{H}_1$ is true. For very small values of missdetection probability a similar thing occurs but in this case the hypothesis $\mathcal{H}_0$ is more weighted in the final sum in equation (77). A special case occurs when $P_{D,l} = P_{F,l}$, $P_0 = P_1$ and $t \to \infty$. In this situation, the scheme converts into the more simple majority decision approach, which just sums all u_l and compares with zero. Even though its simplicity, the maximum likelihood approach performs better than the majority decision scheme in the minimum probability of error criteria as can be seen in figure 14 Filio et al. (2010). The perceptive reader must have noticed by now that there might be some confusion at the fusion centre when there exists an even number of sensors and there is a tie in the decision. This can be settled by considering the a priori probabilities P_0 and P_1 which are inherent to the system.

4. Conclusion

In the first part of this chapter, we have investigated the impact the scattering environment on the performance of primary user detection in multiantennae confinguration. An approximate expressions for the probability of missed detection in function of the number of antennas, parameters of the scattering environment and number of observations. It is shown that at low SNR it is more beneficial to utilize just a single antenna and large number of time samples. This allowes for a better noise suppression via time averaging. On the contrary, at high SNR, it is more beneficial to have more antennas in order to mitigate fading which is a dominant cause of errors in detection in weak noise. It was also shown that for a very strong correlation, *i.e.* $\rho \approx 1$, the equivalent number of antennas is almost unity, $N_{eq} \approx 1$. Therefore this scheme can be usefully applied in a low SNR situations assuming enough time samples are obtained. As $\rho \approx 0$, the diversity gain is increased, therefore making it suitable for higher SNR situations.
The second part of the chapter was devoted to application of the sequential analysis technique to a chive a faster spectrum sensing in cognitive radio networks. It was shown that using the sequential probability ratio test it is possible to detect the presence of a primary user almost twice as fast as other fixed sample approaches such as Neyman Pearson detectors. This can be achieved when dealing in the low SNR region which is a quite often operating characteristic in real life. The effect of error in estimation phase of the carrier on the duration of the sequential analysis has also been investigated. It was shown that the impact of non-coherent detection in sensing the presence of primary users using sequential analysis is the increase of almost twice the samples needed in comparison to a coherent detection approach. Afterwards we derived an optimal fusion rule using detectors that use sequential analysis for taking decisions. We assessed the performance of the system in terms of the time that it takes to gather the decision from all detectors. It was shown that for faster decision, the fusion centre does not consider the opinions of all sensors and hence the performance gets reduced. On the other hand, as

the time of decision increases the performance is better but the system experiment a higher latency.

5. Appendix

Performance derivation of data fusion rule

Let us introduce the following notations

$$a_i = \begin{cases} \ln \frac{1-P_{MD}}{P_{FA}} & \text{if } u_i > 0 \\ \ln \frac{1-P_{FA}}{P_{MD}} = -\ln \frac{P_{MD}}{1-P_{FA}} & \text{if } u_i < 0 \end{cases}, \tag{81}$$

and

$$\xi_i = \begin{cases} a_i = \ln \frac{1-P_{MD}}{P_{FA}} & \text{if } u_i > 0 \\ b_i = -a_i = -\ln \frac{1-P_{FA}}{P_{MD}} & \text{if } u_i < 0 \end{cases}. \tag{82}$$

Consider a $\mathcal{T}$ (test statistic) given by eq. (77)

$$\mathcal{T} = a_0 + \sum_{k=1}^{K} a_k u_k = a_0 + \sum_{k=1}^{K} \xi_k |u_k| = a_0 + \sum_{k=1}^{K} \xi_k. \tag{83}$$

Here ξ_k could be considered as a random variable with PDF

$$\begin{aligned} P_{\xi}(x) = & \quad P_{+}\delta(x-a) + P_{-}\delta x - b \\ & = P\delta(x-a) + (1-P)\delta(x-b), \end{aligned} \tag{84}$$

where

$$\begin{aligned} a = & \ a_i & = \ln \frac{1-P_{MD}}{P_{FA}} \\ b = & -a_i & = -\ln \frac{1-P_{FA}}{P_{MD}} \end{aligned}. \tag{85}$$

and P_{+} is probability of $u = +1$ decision, equal to

$$\begin{aligned} P_{+} = & \quad p(\mathcal{H}_1)(1-P_{MD}) + p(\mathcal{H}_0)P_{FA} \\ & = p(\mathcal{H}_1)(1-P_{MD}) + [1-p(\mathcal{H}_1)]\, P_{FA} \end{aligned}, \tag{86}$$

The corresponding characteristic function of ξ is then given by

$$\Theta_{\xi}(s) = P_{+}e^{-sa} + P_{-}e^{-sb}, \tag{87}$$

and the characteristic function of $\mathcal{T}$ could be evaluated as

$$\begin{aligned} \Theta_{\mathcal{T}} = \Theta_{\xi}^{K} e^{-sa_0} = & \quad \left[P_{+}e^{-sa} + (1-P_{+})e^{-sb}\right]^{K} e^{-sa_0} \\ & = \sum_{k=0}^{K} \binom{K}{k} P_{+}^{k} (1-P_{+})^{K-k} e^{-s[ka+(K-k)b+a_0]} \end{aligned}. \tag{88}$$

Equivalently, the PDF is given by

$$P_{\mathcal{T}}(x) = \sum_{k=0}^{K} \binom{K}{k} P_{+}^{k} (1-P_{+})^{K-k} \delta[x - (ka + (K-k)b + a_0)]. \tag{89}$$

If $k = 0$ then

$$ka + (K-k)b - a_0 = Kb + a_0$$

$$= -K \ln \frac{1 - P_{FA}}{P_{MD}} + \ln \frac{P(\mathcal{H}_1)}{1 - P(\mathcal{H}_1)}. \tag{90}$$

If $P(\mathcal{H}_1) \approx 1$ such that

$$\frac{P(\mathcal{H}_1)}{1 - P(\mathcal{H}_1)} > \left(\frac{1 - P_{FA}}{P_{MD}} \right)^K, \tag{91}$$

then the FC makes only $\mathcal{H}_1$ decisions i.e.

$$P_{MD} = 0, \qquad P_{FA} = P(\mathcal{H}_0) = 1 - p(\mathcal{H}_1).$$

If (91) is not satisfied then there is $k_{\max} > 0$ such that

$$k_{\max} a + (K - k_{\max}) b + a_0 < 0, \tag{92}$$

and

$$(k_{\max} + 1) a + (K - k_{\max} - 1) b + a_0 > 0. \tag{93}$$

In this case the scheme suggested in Chair & Varshney (1986) is equivalent to $(k_{\max} + 1)$ out of K scheme (this is assuming that are statistically equivalent).
Let $\mathcal{H}_1$ be true. Then the target is missed if there are no more than $k_{\max}$ positive decisions, or, equivalently, no less than $K - k_{\max}$ negative decisions. The probability of miss detection at FC is then given by

$$P_{MD_F} = \sum_{k=0}^{k_{\max}} \binom{K}{k} P_{MD}^k (1 - P_{MD})^{K-k}. \tag{94}$$

To more decisions $\mathcal{H}_1$ there should be at least $k_{\max} + 1$ partial 1. If $\mathcal{H}_0$ is true, the probability of false alarm at the fusion center is then:

$$\begin{array}{rl} \mathcal{H}_1 : P_{MD_F} = & \sum_{k=0}^{k_{\max}} \binom{K}{k} (1 - P_{MD})^k P_{MD}^{K-k} \\ \mathcal{H}_0 : \; P_{FA_F} = & \sum_{k=k_{\max}}^{K} \binom{K}{k} (P_{FA})^k (1 - P_{FA})^{K-k} \end{array}. \tag{95}$$

6. References

Abdi, A. & Kaveh, M. (2002). A space-time correlation model for multielement antenna systems in mobile fading channels, 20(3): 550–560.

Akyildiz, I., Lee, W., Vuran, M. & Mohanty, S. (2008). A Survey on Spectrum Management in Cognitive Radio Networks, 46(4): 40–48.

Almalfouh, S. & Stuber, G. (2010). Interference-aware power allocation in cognitive radio networks with imperfect spectrum sensing, *Proc. of ICC 2010*, pp. 1 –6.

Andronov, I. & Fink, L. (1971). *Peredacha diskretnykh soobshchenii po parallel'nym kanalam*, Sov. Radio, Moscow.

Chair, Z. & Varshney, P. (1986). Optimal data fusion in multiple sensor detection systems, 22(1): 98 –101.

Chenhui, H., Xinbing, W., Zichao, Y., Jianfeng, Z., Youyun, X. & Xinbo, G. (2010). A geometry study on the capacity of wireless networks via percolation, 58(10): 2916 – 2925.

Editors, S. (2006). UMTS: Spatial channel model for MIMO simulations, *Tech. report 25.996*.

Federal Communications Commission (November). Spectrum policy task force, *Technical report*, FCC.

Filio, O., Kontorovich, V. & Primak, S. (2011). Characteristics of sequencial detection in cognitive radio networks, *Proc. ICACT 2011*, Phoenix Park, Korea.

Filio, O., Kontorovich, V., Primak, S. & Ramos-Alarcon, F. (2010). Collaborative spectrum sensing for cognitive radio: Diversity combining approach, *Wireless Sensor Network* 3(1): 24–37.

Filio, O., Primak, S. & Kontorovich, V. (2011). On performance of wald test in partially coherent channels, *Proc. ICCIT 2011*, Aqaba, Jordan.

Gosan, N., Jemin, L., Hano, W., Sungtae, K., Sooyong, C. & Daesik, H. (2010). Throughput analysis and optimization of sensing-based cognitive radio systems with markovian traffic, 59(8): 4163–4169.

Haghighi, S. J., Primak, S., Kontorovich, V. & Sejdic, E. (2010). *Mobile and Wireless Communications Physical layer development and implementatiom*, InTech Publishing, chapter Wireless Communications and Multitaper Analysis: Applications to Channel Modelling and Estimation.

Haykin, S. (2005). Cognitive radio: brain-empowered wireless communications, 23(2): 201–220.

Haykin, S., Thomson, D. & Reed, J. (2009). Spectrum sensing for cognitive radio, *Proceedings of the IEEE* 97(5): 849 –877.

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution. Lecture Notes in Statistics. Vol. 9*, Springer-Verlag, New York, Berlin.

Kang, H., Song, I., Yoon, S. & Kim, Y. (2010). A class of spectrum-sensing schemes for cognitive radio under impulsive noise circumstances: Structure and performance in nonfading and fading environments, 59(9): 4322 – 4339.

Kay, S. (1998). *Statistical Signal Processing: Detection Theory*, Prentice Hall PTR, Upper Saddle River, NJ.

Kontorovich, V., Ramos-Alarcón, F., Filio, O. & Primak, S. (2010). Cyclostationary spectrum sensing for cognitive radio and multiantenna systems, *Wireless Communications and Signal Processing (WCSP), 2010 International Conference on*, pp. 1 –6.

Kotz, S. & Adams, J. W. (1964). Distribution of sum of identically distributed exponentially correlated gamma-variables, *The Annals of Mathematical Statistics* 35(1): 277–283.

Li, H. & Han, Z. (2010a). Catch me if you can: An abnormality detection approach for collaborative spectrum sensing in cognitive radio networks, 9(11): 3554 – 3565.

Li, H. & Han, Z. (2010b). Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part i: Known channel statistics, 9(11): 3566 – 3577.

Lin, S.-C., Lee, C.-P. & Su, H.-J. (2010). Cognitive Radio with Partial Channel State Information at the Transmitter, 9(11): 3402–3413.

Middleton (1960). *Introduction to Statistical Communications Theory*, first edn, McGraw-Hill, New York.

Mitola, J. & Maguire, G. (1999). Cognitive radio: making software radios more personal, 6(4): 13–18.

Mitran, P., Le, L. B. & Rosenberg, C. (2010). Queue-aware resource allocation for downlink ofdma cognitive radio networks, 9(10): 3100 – 3111.

Paulraj, A., Nabar, R. & Gore, D. (2003). *Introduction to Space-Time Wireless Communications*, Cambridge University Press, Cambridge, UK.

Poor, H. & Hadjiliadis, O. (2008). *Quickest detection*, Cambridge University Press.

Rui, Z. (2010). On active learning and supervised transmission of spectrum sharing based cognitive radios by exploiting hidden primary radio feedback, 58(10): 2960 – 2970.

Shengli, X., Yi, L., Yan, Z. & Rong, Y. (2010). A parallel cooperative spectrum sensing in cognitive radio networks, 59(8): 4079 – 4092.

Simon, M. & Alouini, M.-S. (2000). *Digital Communication over Fading Channels:A Unified Approach to Performance Analysis*, John Wiley & Sons, New York.

Slepian, D. (1978). Prolate spheroidal wave functions, Fourier analysis, and uncertainty. V-The discrete case, *Bell System Technical Journal* 57: 1371–1430.

Song, M.-G., Kim, D. & Im, G.-H. (2010). Recursive channel estimation method for ofdm-based cooperative systems, 14(11): 1029 – 1031.

Sun, H., Laurenson, D. & Wang, C.-X. (2010). Computationally tractable model of energy detection performance over slow fading channels, 14(10): 924 –926.

Tang, P. K. & Han, C. (2010). On the modeling and performance of three opportunistic spectrum access schemes, 59(8): 4070 – 4078.

Thomson, D. (1982). Spectral estimation and harmonic analysis, 70(9): 1055–1096.

van Trees, H. (2001). *Detection, Estimation, and Modulation Theory: Part I*, first edn, John Wiley & Sons, New York.

von Mises, R. (1964). *Mathematical theory of probability and statistics*, Academic Press, New York.

Wald, A. (2004). *Sequential Analysis*, 6th edn, Dover Phoenix, Mineola, NY.

Wang, R. & Meixia, T. (2010). Blind spectrum sensing by information theoretic criteria for cognitive radios, 59(8): 3806 – 3817.

Wenyi, Z. (2010a). Spectrum shaping: A new perspective on cognitive radioŮpart ii: Coexistence with uncoded legacy transmission, 58(10): 2971 – 2983.

Wenyi, Z. M. (2010b). Spectrum shaping: a new perspective on cognitive radio-part i: coexistence with coded legacy transmission, 58(6): 1857 – 1867.

Yücek, T. & Arslam, H. (2009). A survey of spectrum sensing algorithms for cognitive radio applications, *IEEE Comm. Surveys & Tutorials* 11(1): 116–130.

Zhang, R., Lim, T. J., Liang, Y.-C. & Zeng, Y. (2010). Multi-antenna based spectrum sensing for cognitive radios: A GLRT approach, 58(1): 84–88.

Zou, Q., Zheng, S. & Sayed, A. (2010). Cooperative sensing via sequential detection, 58(12): 6266 –6283.

8

Multi-Cell Cooperation for Future Wireless Systems

A. Silva, R. Holakouei and A. Gameiro
University of Aveiro/Instituto de Telecomunicações (IT)
Portugal

1. Introduction

The wireless communications field is experiencing a rapid and steady growth. It is expected that the demand for wireless services will continue to increase in the near and medium term, asking for more capacity and putting more pressure on the usage of radio resources. The conventional cellular architecture considers co-located multiple input multiple output (MIMO) technology, which is a very promising technique to mitigate the channel fading and to increase the cellular system capacity (Foschini & Gans, 1998). On the other hand, orthogonal frequency division multiplexing (OFDM) is a simple technique to mitigate the effects of inter-symbol interference in frequency selective channels (Uppala & Li, 2004), (Bahai et al., 2004). However, the problems inherent to these systems such as shadowing, significant correlation between channels in some environments and intercell interference significantly degrade the capacity gains promised by MIMO techniques (Andrews et al., 2007). Although theoretically attractive, the deployment of MIMO in commercial cellular systems is limited by interference between neighbouring cells, and the entire network is essentially interference-limited (Foschini et al., 2006; Mudumbai et al., 2009).

Conventional approaches to mitigate multi-cell interference, such as static frequency reuse and sectoring, are not efficient for MIMO-OFDM networks as each has important drawbacks (Andrews et al., 2007). Universal frequency reuse (UFR), meaning that all cells/sectors operate on the same frequency channel, is mandatory if we would like to achieve spectrally-efficient communications. However, as it is pointed out in (Foschini et al., 2006), this requires joint optimization of resources in all cells simultaneously to boost system performance and to reduce the radiated power. Such systems have the advantage of macro-diversity that is inherent to the widely spaced antennas and more flexibility to deal with intercell interference, which fundamentally limits the performance of user terminals (UTs) at cell edges (Andrews et al., 2007). Different transmit strategies can be considered, depending on the capacity of the backhaul channel that connects the coordinated base stations. Recently, an enhanced cellular architecture with a high-speed backhaul channel has been proposed and implemented, under the European FUTON project (FUTON, 2011), (Diehm et al., 2010). This project aims at the design of a distributed broadband wireless system (DBWS) by carrying out the development of a radio over fiber (RoF) infrastructure transparently connecting the BSs to a central unit (CU) where centralized joint processing can be performed. Also, multi-cell cooperation is already under study in LTE under the Coordinated

Multipoint (CoMP) concept (3GPP LTE, 2007) that although not included in the current releases, will probably be specified for the future ones.

In recent years, relevant works on multi-cell precoding techniques have been proposed in (Jing et al., 2008), (Somekh et al, 2007), (Boccardi & Huang, 2007), (Zhang et al, 2009), (Marsch & Fettweis, 2009), (Armada et al., 2009), (Kobayashi et al., 2009), (Zhang, 2010), (Bjornson et al., 2010). The multi-cell downlink channel is closely related to the MIMO broadcast channel (BC), where the optimal precoding is achieved by the dirty paper coding (DPC) principle (Costa, 1983). However, the significant amount of processing complexity required by DPC prohibits its implementation in practical multi-cell processing. Some suboptimal multi-cell linear precoding schemes have been discussed in (Jing et al., 2008), where analytical performance expressions for each scheme were derived considering nonfading scenario with random phases. The comparison of the achievable rates by the different proposed cooperative schemes showed a tradeoff between performance improvement and the requirement for BS cooperation, signal processing complexity and channel state information (CSI) knowledge. In (Somekh et al, 2007) the impact of joint multi-cellsite processing was discussed through a simple analytically tractable circular multi-cell model. The potential improvement in downlink throughput of cellular systems using limited network coordination to mitigate intercell interference has been discussed in (Boccardi & Huang, 2007), where zero forcing (ZF) and DPC precoding techniques under distributed and centralized architectures have been studied. In (Zhang et al, 2009) a clustered BS coordination is enabled through a multi-cellblock diagonalization (BD) scheme to mitigate the effects of interference in multi-cell MIMO systems. Three different power allocation algorithms were proposed with different constraints to maximize the sum rate. A centralized precoder design and power allocation was considered. In (Marsch & Fettweis, 2009), the inner bounds on capacity regions for downlink transmission were derived with or without BS cooperation and under per-antenna power or sum-power constraint. The authors showed that under imperfect CSI, significant gains are achievable by BS cooperation using linear precoding. Furthermore the type of cooperation depends on channel conditions in order to optimize the rate/backhaul tradeoff. Two multi-cell precoding schemes based on the waterfilling technique have been proposed in (Armada et al., 2009). It was shown that these techniques achieve a performance, in terms of weighted sum rate, very close to the optimal. In (Kobayashi et al., 2009), each BS performs ZF locally to remove the channel interference and based on the statistical knowledge of the channels, the CU performs a centralized power allocation that jointly minimizes the outage probability of the UTs. A new BD cooperative multi-cells scheme has been proposed in (Zhang, 2010), to maximize the weighted sum-rate achievable for all the UTs. Multiuser multi-cell precoding with distributed power allocation has been discussed in (Bjornson et al., 2010). It is assumed that each BS has only the knowledge of local CSI and based on that the beamforming vectors used to achieve the outer boundary of the achievable rate region was derived considering both instantaneous and statistical CSI. An overview of the theory for multi-cell cooperation in networks has been presented in (Gesbert et al., 2010).

In this chapter we design and evaluate linear precoding techniques for multi-cell MIMO-OFDM cooperative systems. Two approaches are considered: centralized with a high-speed backhaul channel, where it is assumed that full CSI and data are available at the CU; and distributed with lower speed backhaul channel, where only some channel information and data are shared by the BSs. The precoder design aims at two goals: allow spatial users separation and optimize the power allocation. The two problems can be decoupled leading to a two step design: the precoder vectors design and power allocation algorithms. In this chapter we discuss three centralized power allocation algorithms with different complexities and per-BS power constraint: one optimal to minimize the average bit error rate (BER), for which the powers can be obtained numerically by using convex optimization, and two suboptimal. In this latter approach, the powers are computed in two phases. First the powers are derived under total power constraint (TPC). Two criterions are considered, namely minimization of the average BER, which leads to an iterative approach and minimization of the sum of inverse of signal-to-noise ratio for which closed form solution is achieved. Then, the final powers are computed to satisfy the individual per-BS power constraint.

The rest of this chapter is organized as follows: in section 2 the general scenario is described, section 3 discusses centralized multi-cell MIMO OFDM cooperative precoding schemes, while in section 4 distributed multi-cell MIMO OFDM cooperative schemes are proposed, in section 5 the simulation results are presented and discussed. Finally, conclusions are drawn in section 6.

Throughout this chapter, we will use the following notations. Lowercase letters, boldface lowercase letters and boldface uppercase letters are used for scalars, vectors and matrices, respectively. $(.)^H, (.)^T, (.)^*$ represent the conjugate transpose, the transpose and complex conjugate operators, respectively. $\mathrm{E}[.]$ represents the expectation operator, $\mathbf{I}_N$ is the identity matrix of size $N \times N$, $\mathcal{CN}(.,.)$ denotes a circular symmetric complex Gaussian vector, $[\mathbf{A}]_{i,j}$ is the *(i,j)*th element and $[\mathbf{A}]_i$ is the *i*th column of the matrix $\mathbf{A}$.

2. Scenario description

Multi-cell architectures that assume a global coordination can eliminate the intercell interference completely. However, in practical cellular scenarios, issues such as the complexity of joint signal processing of all the BSs, the difficulty in acquiring full CSI from all UTs at each BS, and synchronization requirements will make global coordination difficult. Therefore, in this chapter we assume a clustered multi-cell cellular system as shown in Fig. 1, where the BSs are linked to a central unit (e.g., by optical fiber) as proposed in (FUTON, 2010). In such architecture the area covered by the set of cooperating BSs is termed as super-cell. The area defined by all the super-cells that are linked to the same CU is termed as serving area. The BSs corresponding to a super-cell are processed jointly by a joint processing unit (JPU). The number of cooperating BSs per super-cell should not be high for the reasons discussed above. In this chapter, it is assumed that the interference between the super-cells is negligible. In fact as we are replacing the concept of cell by the one of super-cell, this means that there will be some interference among the super-cells especially at the edges. Two approaches can be considered to deal with the inter-super-cell interference. The precoders are designed to remove both intra-super-cell and inter-super-cell interference, but as discussed in (Somekh et al., 2007) this strategy reduces the number of degrees of freedom to efficiently eliminate the intra-super-cell interference. Alternatively, the radio resource management can be jointly performed for a large set of super-cells (the serving area) at the CU, and thus the resource allocation can be done in a way that the UTs of each super-cell edge interfere as little as possible with the users of other super-cells (FUTON, 2010), justifying our assumption to neglect it. This resource allocation problem is however beyond the scope of this chapter. In this latter approach all degrees of freedom can be used to efficiently eliminate the intra-super-cell interference.

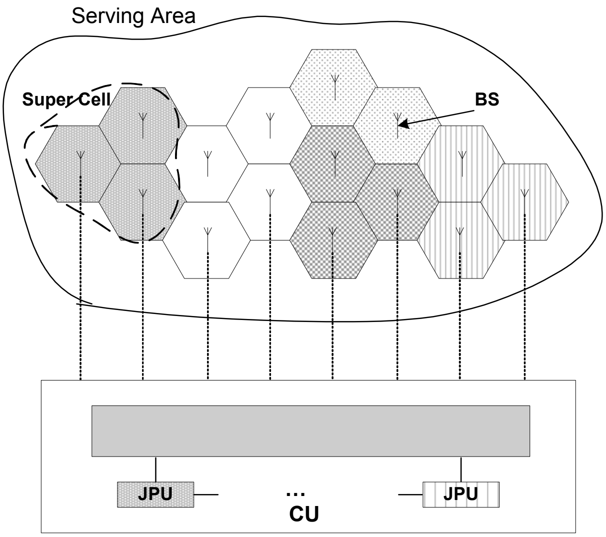


Fig. 1. Enhanced cellular architecture

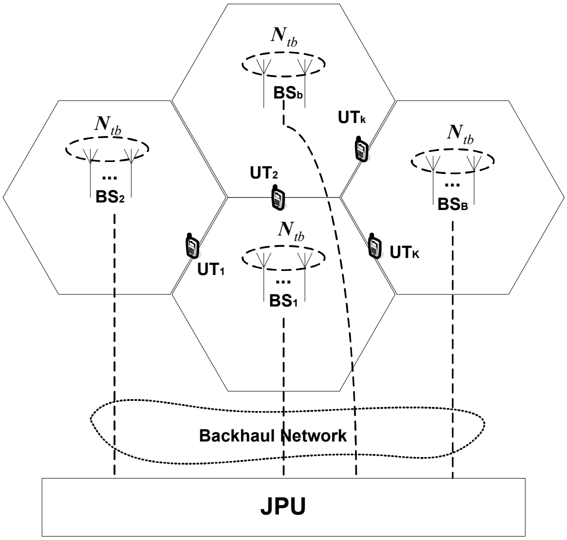


Fig. 2. Multi-cell system overview

We consider a scenario of B BSs comprising a super-cell; each BS is equipped with N_{t_b} antennas, transmitting to K UTs as shown in Fig. 2. The total number of transmitting antennas per-super-cell is N_t. User k is equipped with single antenna or an antenna array of N_{r_k} elements and the total number of receiving antennas per-super-cell is N_r, which is equal to the number of users K in case of single antenna UTs. Also, we assume an OFDM based system with N_c parallel frequency flat fading channels.

3. Centralized multi-cell based system

We consider a multi-cell system based on the scenario defined in previous section where the BSs are transparently linked by optical fiber to a central unit. Thanks to the high speed backhaul, we can assume that all the information of all BSs, i.e., full CSI and data, belonging to the same super-cell are available at the JPU. Thus, to remove the multi-cell multiuser interference we can use a similar linear precoding algorithm designed for single cell based systems. The major difference between multi-cell and single cell systems is that the power constraints have to be considered on a per-BS basis instead. The proposed schemes are considered in two phases: singular value decomposition SVD based precoding and power allocation.

3.1 System model

To build up the mathematical model we consider that user $k, k=1,...,K$ can receive up to N_{r_k} data symbols on subcarrier $l, l=1,...,N_c$ i.e., $\boldsymbol{x}_{k,l}=[x_{k,1,l} \ \cdots \ x_{k,N_{r_k},l}]^T$ and the global symbol vector, comprising all user symbol vectors, is $\boldsymbol{x}_l=[\boldsymbol{x}_{1,l}^T \ \ldots \ \boldsymbol{x}_{K,l}^T]^T$ of size $N_r \times 1$.

The data symbol of user *k* on subcarrier *l*, is processed by the transmit precoder $\mathbf{W}_{k,l} \in C^{N_t \times N_{r_k}}$ in JPU, before being transmitted over BSs antennas. These individual precoders together form the global transmit precoder matrix on subcarrier l, $\mathbf{W}_l=\begin{bmatrix}\mathbf{W}_{1,l} & \cdots & \mathbf{W}_{K,l}\end{bmatrix}$ of size $N_t \times N_r$. Let the downlink transmit power over the N_t distributed transmit antennas for user *k* and data symbol $i, i=1,...,N_{r_k}$ on subcarrier *l*, be $p_{k,i,l}$, with $\mathbf{p}_{k,l}=\begin{bmatrix}p_{k,1,l} & \cdots & p_{k,N_{r_k},l}\end{bmatrix}$ and the global power matrix $\mathbf{P}_l=\text{diag}\{\begin{bmatrix}\mathbf{p}_{1,l} & \cdots & \mathbf{p}_{K,l}\end{bmatrix}\}$ is of size $N_r \times N_r$.
Under the assumption of linear precoding, the signal transmitted by the JPU on subcarrier *l* is given by $z_l=\mathbf{W}_l\mathbf{P}_l^{1/2}\mathbf{x}_l$ and the global received signal vector on subcarrier *l* can be expressed by,

$$\mathbf{y}_l=\mathbf{H}_l\mathbf{W}_l\mathbf{P}_l^{1/2}\mathbf{x}_l+\mathbf{n}_l \tag{1}$$

where $\mathbf{H}_l=\begin{bmatrix}\mathbf{H}_{1,l}^T & \cdots & \mathbf{H}_{K,l}^T\end{bmatrix}^T$ of size $N_r \times N_t$ is the global frequency flat fading MIMO channel on subcarrier l. The channel of user k is represented by $\mathbf{H}_{k,l}=\begin{bmatrix}\mathbf{H}_{1,k,l} & \cdots & \mathbf{H}_{b,k,l} & \cdots & \mathbf{H}_{B,k,l}\end{bmatrix}$ of size $N_{r_k} \times N_t$, and $\mathbf{H}_{b,k,l}$ of size $N_{r_k} \times N_{t_b}$ represents the channel between user k and BS $b, b=1,...,B$ on subcarrier l. The channel $\mathbf{H}_{b,k,l}$ can be decomposed as the product of the fast fading $\mathbf{H}_{b,k,l}^c$ and slow fading $\sqrt{\rho_{b,k}}$ components, i.e., $\mathbf{H}_{b,k,l}=\mathbf{H}_{b,k,l}^c\sqrt{\rho_{b,k}}$, where $\rho_{b,k}$ represents the long-term power gain between BS b and user k and $\mathbf{H}_{b,k,l}^c$ contains the fast fading coefficients with $\mathcal{CN}(0,1)$ entries. $\mathbf{n}_l=\begin{bmatrix}\mathbf{n}_{1,l}^T & \cdots & \mathbf{n}_{K,l}^T\end{bmatrix}^T$ represents the global additive white Gaussian noise (AWGN) vector and $\mathbf{n}_{k,l}=\begin{bmatrix}n_{k,1,l} & \ldots & n_{k,N_{r_k},l}\end{bmatrix}^T$ is the noise at the user k terminal on subcarrier l with zero mean and power σ^2, i.e., $\text{E}[\mathbf{n}_{k,l}\mathbf{n}_{k,l}^H]=\sigma^2\mathbf{I}_{N_{r_k}}$.

The signal transmitted by the BS *b* on subcarrier l can be written as $z_{b,l}=\mathbf{W}_{b,l}\mathbf{P}_l^{1/2}\mathbf{x}_l$, where $\mathbf{W}_{b,l}$ of size $N_{t_b} \times N_r$ represents the global precoder at BS *b* on subcarrier l. The average transmit power of BS *b* is then given by,

$$\mathrm{E}\left[\left\|\mathbf{z}_b\right\|^2\right]=\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c} p_{k,i,l}\ \left[\mathbf{W}_{b,k,l}^{H}\mathbf{W}_{b,k,l}\right]_{i,i} \tag{2}$$

where $\mathbf{z}_b$ is the signal transmitted over the N_c subcarriers and $\mathbf{W}_{b,k,l}$ of size $N_{t_b}\times N_{r_k}$ represents the precoder of user k on subcarrier l at BS b.

3.2 Centralized precoder vectors

In this section, we consider the SVD based precoding algorithm similar to the one proposed in (Yu et al., 2004). We assume that $N_t \geq N_r$. Briefly, we define $\tilde{\mathbf{H}}_{k,l}$ as the following $\left(N_r - N_{r_k}\right)\times N_t$ matrix,

$$\tilde{\mathbf{H}}_{k,l}=\left[\mathbf{H}_{1,l}...\mathbf{H}_{k-1,l},\mathbf{H}_{k+1,l}...\mathbf{H}_{K,l}\right]^T \tag{3}$$

If we denote rank of $\tilde{\mathbf{H}}_{k,l}$ as $\tilde{L}_{k,l}$ then the null space of $\tilde{\mathbf{H}}_{k,l}$ has dimension of $N_t - \tilde{L}_{k,l} \geq N_{r_k}$. The SVD of $\tilde{\mathbf{H}}_{k,l}$ is partitioned as follows,

$$\tilde{\mathbf{H}}_{k,l}=\tilde{\mathbf{U}}_{k,l}\tilde{\mathbf{D}}_{k,l}\left[\tilde{\mathbf{V}}_{k,l}^{(0)}\,\tilde{\mathbf{V}}_{k,l}^{(1)}\right]^H \tag{4}$$

where $\tilde{\mathbf{V}}_{k,l}^{(0)}$ holds the $N_t - \tilde{L}_{k,l}$ singular vectors in the null space of $\tilde{\mathbf{H}}_{k,l}$. The columns of $\tilde{\mathbf{V}}_{k,l}^{(0)}$ are candidate for user k precoding matrix $\mathbf{W}_{k,l}$, causing zero gain at the other users, hence result in an effective SU-MIMO system. Since $\tilde{\mathbf{V}}_{k,l}^{(0)}$ potentially holds more precoders than the number of data streams user k can support, an optimal linear combination of these vectors must be found to build matrix $\mathbf{W}_{k,l}$, which can have at most N_{r_k} columns. To do this, the following SVD is formed,

$$\mathbf{H}_{k,l}\tilde{\mathbf{V}}_{k,l}^{(0)}=\mathbf{U}_{k,l}\mathbf{D}_{k,l}\left[\mathbf{V}_{k,l}^{(0)}\,\mathbf{V}_{k,l}^{(1)}\right]^H \tag{5}$$

where $\mathbf{D}_{k,l}$ is $L_{k,l}\times L_{k,l}$ and $\mathbf{V}_{k,l}^{(1)}$ represents the $L_{k,l}$ singular vectors with non-zero singular values. The $L_{k,l} \leq N_{r_k}$ columns of the product $\tilde{\mathbf{V}}_{k,l}^{(0)}\mathbf{V}_{k,l}^{(1)}$ represent precoders that further improve the performance subject to producing zero inter-user interference. The transmit precoder matrix will thus have the following form,

$$\bar{\mathbf{W}}_l=\left[\tilde{\mathbf{V}}_{1,l}^{(0)}\mathbf{V}_{1,l}^{(1)}\quad ...\quad \tilde{\mathbf{V}}_{K,l}^{(0)}\mathbf{V}_{K,l}^{(1)}\right]\mathbf{P}_l^{1/2}=\mathbf{W}_l\mathbf{P}_l^{1/2} \tag{6}$$

The global precoder matrix with power allocation, $\bar{\mathbf{W}}_l=\left[\mathbf{W}_{1,l}\ \ ...\ \ \mathbf{W}_{K,l}\right]\mathbf{P}_l^{1/2}$ as computed above, block-diagonalizes the global equivalent channel $\mathbf{H}_l$, i.e., $\mathbf{H}_l\bar{\mathbf{W}}_l=\mathrm{diag}\left\{\left[\mathbf{H}_{e,1,l},\ldots,\mathbf{H}_{e,K,l}\right]\right\}$ and the interference is completely removed considering perfect CSI.

Let us define $\mathbf{H}_{e,k,l}$=$\mathbf{H}_{k,l}\bar{\mathbf{W}}_{k,l}=\mathbf{H}_{k,l}\mathbf{W}_{k,l}\mathbf{P}_{k,l}^{1/2}$ of size $N_{r_k}\times N_{r_k}$ as the equivalent enhanced channel for user k on subcarrier l, where $\mathbf{P}_{k,l}$=diag$\{\mathbf{p}_{k,l}\}$ is of size $N_{r_k}\times N_{r_k}$. Rewriting equation (1) for this user, we have,

$$\mathbf{y}_{k,l}=\mathbf{H}_{e,k,l}\mathbf{x}_{k,l}+\mathbf{n}_{k,l} \tag{7}$$

To estimate $\mathbf{x}_{k,l}$, user k processes $\mathbf{y}_{k,l}$ by doing maximal ratio combining (MRC), and the soft decision variable $\hat{\mathbf{x}}_{k,l}$ is given by

$$\hat{\mathbf{x}}_{k,l}=\mathbf{H}_{e,k,l}^{H}\mathbf{y}_{k,l}=\mathbf{H}_{e,k,l}^{H}\mathbf{H}_{e,k,l}\mathbf{x}_{k,l}+\mathbf{H}_{e,k,l}^{H}\mathbf{n}_{k,l} \tag{8}$$

It should be mentioned that channel $\mathbf{H}_{e,k,l}$ can be easily estimated at UT k . It can be shown that,

$$\mathbf{H}_{e,k,l}^{H}\mathbf{H}_{e,k,l}=\operatorname{diag}\left\{\left[p_{k,1,l}\lambda_{k,1,l},\ldots,p_{k,N_{r,k},l}\lambda_{k,N_{r_k},l}\right]\right\} \tag{9}$$

where $\sqrt{\lambda_{k,i,l}}$ is the i^{th} singular value of matrix $\mathbf{H}_{k,l}\mathbf{W}_{k,l}$. From equations (8) and (9) is easy to see that the instantaneous SNR of data symbol i of user k on subcarrier l can be written as

$$\mathrm{SNR}_{k,i,l}=\frac{p_{k,i,l}\lambda_{k,i,l}}{\sigma^2} \tag{10}$$

From (10), assuming a M-ary QAM constellations, the instantaneous probability of error of data symbol i of user k on subcarrier l is given by (Proakis, 1995),

$$P_{e,k,i,l}=\psi Q\left(\sqrt{\beta SNR_{k,i,l}}\right) \tag{11}$$

where $Q(x)=\left(1/\sqrt{2\pi}\right)\int_{x}^{\infty}e^{-\left(t^2/2\right)}dt$, $\beta=3/\left(M-1\right)$ and $\psi=\left(4/\log_2 M\right)\left(1-1/\sqrt{M}\right)$.

3.3 Power allocation strategies

Once the multi-cell multiuser interference removed, the power loading elements of $\mathbf{P}_l$ can be computed in order to minimize or maximize some metrics. Most of the proposed power allocation algorithms for precoded multi-cell based systems have been designed to maximize the sum rate, e.g., (Jing et al., 2008; Bjornson et al., 2010). In this paper, the criteria used to design power allocation are minimization of the average BER and sum of inverse of SNRs, which essentially lead to a redistribution of powers among users and therefore provide users fairness (which in practical cellular systems may be for the operators a goal as important as throughput maximization). The aim of these power allocation schemes is to improve the user's fairness, namely inside each super-cell.

A. Optimal minimum BER power allocation

We minimize the instantaneous average probability under the per-BS power constraint P_{tb}, i.e., $\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}\,\left[\mathbf{W}_{b,k,l}^{H}\mathbf{W}_{b,k,l}\right]_{i,i}\leq P_{tb},\ b=1,\ldots,B$. Without loss of generality, we assume a 4-QAM constellation, and thus the optimal power allocation problem with per-BS power constraint can be formulated as,

$$\min_{\{p_{k,i,l}\}}\left(\frac{1}{KN_{r_k}N_c}\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}Q\left(\sqrt{\frac{p_{k,i,l}\lambda_{k,i,l}}{\sigma^2}}\right)\right)\ \text{s.t.}\ \begin{cases}\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}\,\left[\mathbf{W}_{b,k,l}^{H}\mathbf{W}_{b,k,l}\right]_{i,i}\leq P_{tb}\ ,b=1,..,B\\ p_{k,i,l}\geq 0,\ k=1,..,K,i=1,..,N_{r_k},l=1,..,N_c\end{cases} \tag{12}$$

Since the objective function is convex in $p_{k,i,l}$, and the constraint functions are linear, this is a convex optimization problem. Therefore, it may be solved numerically by using for example the interior-point method (Boyd & Vandenberghe, 2004). This scheme is referred as centralized per-BS optimal power allocation (Cent. per-BS OPA).

B. Suboptimal power allocation approaches

Since the complexity of the above scheme is too high, and thus it could not be of interest for real wireless systems, we also resort to less complex suboptimal solutions. The proposed strategy has two phases: first the power allocation is computed by assuming that all BSs of each super-cell can jointly pool their power, i.e., a TPC P_t is imposed instead and the above optimization problem reduces to,

$$\min_{\{p_{k,i,l}\}}\left(\frac{1}{KN_{r_k}N_c}\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}Q\left(\sqrt{\frac{p_{k,i,l}\lambda_{k,i,l}}{\sigma^2}}\right)\right)\ \text{s.t.}\begin{cases}\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}\ \left[\mathbf{W}_{k,l}^{H}\mathbf{W}_{k,l}\right]_{i,i}\le P_t\\ p_{k,i,l}\ge 0,\ k=1,..,K,i=1,..,N_{r_k},l=1,..,N_c\end{cases}\tag{13}$$

with $\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}\ \left[\mathbf{W}_{k,l}^{H}\mathbf{W}_{k,l}\right]_{i,i}=\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}$, note that the N_{r_k} columns of $\mathbf{W}_{k,l}$ have unit norm. Using the Lagrange multipliers method (Haykin, 1996), the following cost function with μ Lagrange multiplier is minimized,

$$J_{c,1}=\frac{1}{KN_{r_k}N_c}\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}Q\left(\sqrt{\frac{p_{k,i,l}\lambda_{k,i,l}}{\sigma^2}}\right)+\mu\left(\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}-P_t\right)\tag{14}$$

The powers $p_{k,i,l}$ can be determined by setting the partial derivatives of $J_{c,1}$ to zero and as shown in (Holakouei et al., 2011), the solution is

$$p_{k,i,l}=\frac{\sigma^2}{\lambda_{k,i,l}}W_0\left(\frac{\lambda_{k,i,l}^2}{8\pi\mu^2\left(KN_{r_k}N_c\right)^2\sigma^4}\right)\tag{15}$$

where W_0 stands for Lambert's W function of index 0 (Corless et al., 1996). This function $W_0(x)$ is an increasing function. It is positive for $x>0$, and $W_0(0)=0$. Therefore, μ^2 can be determined iteratively to satisfy $\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}=P_t$. The optimization problem of (13) is similar to the single cell power allocation optimization problem, where the users are allocated the same total multi-cell power, which may serve as a lower bound of the average BER for the multi-cell with per-BS power constraint. One solution based on Lambert W function that minimizes the instantaneous BER was also derived in the context of single user single cell MIMO systems (Rostaing et al., 2002).

The second phase consists in scaling the power allocation matrix $\mathbf{P}_l$ by a factor of β in order to satisfy the individual per-BS power constraints as discussed in (Zhang et al., 2009) which can be given by

$$\beta=\frac{P_{tb}}{\max_{b=1,\dots,B}\left(\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}\ \left[\mathbf{W}_{b,k,l}^{H}\mathbf{W}_{b,k,l}\right]_{i,i}\right)}\tag{16}$$

This scaled power factor assures that the transmit power per-BS is less or equal to P_{tb}. Note that this factor is less than one and thus the SNR given by (10) has a penalty of $10\log(\beta)$ dB. This scheme is referred as centralized per-BS suboptimal iterative power allocation (Cent. per-BS SOIPA).

Although this suboptimal solution significantly reduces the complexity relative to the optimal one, it still needs an iterative search. To further simplify we propose an alternative power allocation method based on minimizing the sum of inverse of SNRs, and a closed-form expression can be obtained. Note that minimizing the sum of inverse of SNRs is similar to the maximization of the harmonic mean of the SINRs discussed in (Palomar, 2003). In this case, the optimization problem is written as,

$$\min_{\{p_{k,i,l}\}}\left(\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}\frac{\sigma^2}{p_{k,i,l}\lambda_{k,i,l}}\right) \text{ s.t. } \begin{cases}\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}\left[\mathbf{W}_{k,l}^{H}\mathbf{W}_{k,l}\right]_{i,i}\leq P_t\\ p_{k,i,l}\geq 0,\ k=1,..,K, i=1,..,N_{r_k}, l=1,..,N_c\end{cases} \tag{17}$$

Since the objective function is convex in $p_{k,i,l}$, and the constraint functions are linear, (17) is also a convex optimization problem. To solve it we follow the same suboptimal two phases approach as for the first problem. First, we impose a total power constraint and the following cost function, using again the Lagrangian multipliers method, is minimized,

$$J_{c,2}=\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}\frac{\sigma^2}{p_{k,i,l}\lambda_{k,i,l}}+\mu\left(\sum_{k=1}^{K}\sum_{i=1}^{N_{r_k}}\sum_{l=1}^{N_c}p_{k,i,l}-P_t\right) \tag{18}$$

Now, setting the partial derivatives of $J_{c,2}$ to zero and after some mathematical manipulations, the powers $p_{k,i,l}$ are given by,

$$p_{k,i,l}=\frac{P_t}{\sqrt{\lambda_{k,i,l}}\sum_{j=1}^{K}\sum_{n=1}^{N_{r_k}}\sum_{p=1}^{N_c}\frac{1}{\sqrt{\lambda_{j,n,p}}}} \tag{19}$$

The second phase consists in scaling the power allocation matrix $\mathbf{P}_l$ by a factor of β, using (19) instead of (15), in order to satisfy the individual per-BS power constraints. This scheme is referred as centralized per-BS suboptimal closed-form power allocation (Cent. per-BS SOCPA).

The above power allocation schemes can also be used, under minor modifications, for the case where the system is designed to achieve diversity gain instead of multiplexing gain. In diversity mode the same user data symbol is received on each receiver antenna, increasing the diversity order. Thus $x_{k,i,l}=x_{k,N_{r_k},l}$, $i=1...N_{r_k}-1$ and then the SNR is given by

$$\text{SNR}_{k,l}=\frac{p_{k,l}\sum_{i=1}^{N_{r_k}}\lambda_{k,i,l}}{\sigma^2}=\frac{p_{k,l}\alpha_{k,l}}{\sigma^2} \tag{20}$$

and the power loading coefficient is computed only per user and subcarrier. In this case to compute the power allocation coefficients we should replace $\lambda_{k,i,l}$ by $\alpha_{k,l}$ and remove the script i in all equations.

4. Distributed multi-cell based system

As discussed in section 2 due to limitations in terms of delay and capacity on backhaul network, it is necessary to reduce signalling overhead. For this purpose, in this section the precoders are designed in a distributed fashion, i.e., based on local CSI at each BS but we still consider data sharing and centralized power allocation techniques.

4.1 System model

Assuming single antennas UTs and under the assumption of linear precoding, the signal transmitted by the BS *b* on sub-carrier *l* is given by,

$$\mathbf{x}_{b,l} = \sum_{k=1}^{K} \sqrt{p_{b,k,l}}\,\boldsymbol{w}_{b,k,l} \mathrm{s}_{k,l}, \tag{21}$$

where $p_{b,k,l}$ represents the power allocated to UT *k* on sub-carrier *l* and BS *b*, $\boldsymbol{w}_{b,k,l} \in \mathbb{C}^{N_{t_b} \times 1}$ is the precoder of user k at BS b on sub-carrier l with unit norms, i.e., $\left\|\boldsymbol{w}_{b,k,l}\right\| = 1,\ b = 1,\ldots,B,\ k = 1,\ldots,K, l = 1,\ldots,N_c$. The data symbol $\mathrm{s}_{k,l}$, with $\mathrm{E}\left[\left|\mathrm{s}_{k,l}\right|^2\right] = 1$, is intended for UT k and is assumed to be available at all BSs. The average power transmitted by the BS b is then given by,

$$\mathrm{E}\left[\left\|\mathbf{x}_b\right\|^2\right] = \sum_{l=1}^{N_c} \sum_{k=1}^{K} p_{b,k,l} \tag{22}$$

where $\mathbf{x}_b$ is the signal transmitted over the N_c subcarriers. The received signal at the UT k on sub-carrier l, $y_{k,l} \in \mathbb{C}^{1\times 1}$, can be expressed by,

$$y_{k,l} = \sum_{b=1}^{B} \boldsymbol{h}_{b,k,l}^{H} \boldsymbol{x}_{b,l} + n_{k,l} \tag{23}$$

where $\boldsymbol{h}_{b,k,l} \in \mathbb{C}^{N_{t_b} \times 1}$ represents the frequency flat fading channel between BS b and UT k on sub-carrier l and $n_{k,l} \sim \mathcal{CN}\left(0,\sigma^2\right)$ is the noise.

The channel $\boldsymbol{h}_{b,k,l}$, as for the centralized approach, can be decomposed as the product of the fast fading $\boldsymbol{h}_{b,k,l}^{c}$ and slow fading $\sqrt{\rho_{b,k}}$ components, i.e., $\boldsymbol{h}_{b,k,l}$=$\boldsymbol{h}_{b,k,l}^{c}\sqrt{\rho_{b,k}}$, where $\rho_{b,k}$ represents the long-term power gain between BS b and user k and $\boldsymbol{h}_{b,k,l}^{c}$ contains the fast fading coefficients with $\mathcal{CN}\left(0,1\right)$ entries. The antenna channels from BS b to user k , i.e. the components of $\boldsymbol{h}_{b,k,l}^{c}$, may be correlated but the links seen from different BSs to a given UT are assumed to be uncorrelated as the BSs of one super-cell are geographically separated.

4.2 Distributed precoder vectors

As discussed above, to design the distributed precoder vector we assume that the BSs have only knowledge of local CSI, i.e., BS b knows the instantaneous channel vectors $\boldsymbol{h}_{b,k,l}, \forall k,l$, reducing the feedback load over the backhaul network as compared with the full centralized precoding approach. We consider a zero forcing transmission scheme with the phase of the received signal at each UT aligned. From (21) and (23) the received signal at UT k on sub-carrier l can be decomposed in,

$$y_{k,l} = \underbrace{\sum_{b=1}^{B} \sqrt{p_{b,k,l}} \boldsymbol{h}_{b,k,l}^H \boldsymbol{w}_{b,k,l} s_{k,l}}_{Desired\ Signal} + \underbrace{\sum_{b=1}^{B} \boldsymbol{h}_{b,k,l}^H \sum_{j=1, j\neq k}^{K} \sqrt{p_{b,j,l}} \boldsymbol{w}_{b,j,l} s_{j,l}}_{Multiuser\ Multicell\ Interference} + \underbrace{n_{k,l}}_{Noise} \tag{24}$$

where $\boldsymbol{w}_{b,k,l}$ is a unit-norm zero forcing vector orthogonal to $K-1$ channel vectors, $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$. Such precoding vectors always exist because we assume that the number of antennas at each BS is higher or equal to the number of single antenna UTs, i.e. $N_{t_b} \geq K$. Note that here *K* is the number of users that share the same set of resources. Considering an OFDMA based system, the total number of users can be significantly larger than *K*, since different set of resources can be shared by different set of users. By using such precoding vectors, the multi-cell interference is cancelled and each data symbol on each subcarrier is only transmitted to its intended UT. Also, for any precoding vector $\bar{\boldsymbol{w}}_{b,k,l}$ in the null space of $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$, $\boldsymbol{w}_{b,k,l} = \bar{\boldsymbol{w}}_{b,k,l} e^{j\varphi}$ is also in the null space of $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$. Thus, we can choose the precoding vectors such that the terms $\boldsymbol{h}_{b,k,l}^H \boldsymbol{w}_{b,k,l}$ all have zeros phases, i.e., $\angle(\boldsymbol{h}_{b,k,l}^H \boldsymbol{w}_{b,k,l}) = 0,\ \forall(b,k,l)$. These precoding vectors can be easily computed, so if $\bar{\mathbf{W}}_{b,k,l}$ is found to lie in the null space of $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$, the final precoding vector $\boldsymbol{w}_{b,k,l},\ b = 1,\ldots,B,$ $k = 1,\ldots,K,\ l = 1,\ldots,N_c$, with the phase of the received signal at each UT aligned, is given by,

$$\boldsymbol{w}_{b,k,l} = \bar{\mathbf{W}}_{b,k,l} \frac{\left(\boldsymbol{h}_{b,k,l}^H \bar{\mathbf{W}}_{b,k,l}\right)^H}{\left\|\boldsymbol{h}_{b,k,l}^H \bar{\mathbf{W}}_{b,k,l}\right\|} \tag{25}$$

where $\bar{\mathbf{W}}_{b,k,l} \in \mathbb{C}^{N_{t_b} \times \left(N_{t_b} - K + 1\right)}$ holds the $\left(N_{t_b} - K + 1\right)$ singular vectors in the null space of $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$. For the case where $N_{t_b} = K$, only one vector lies in the null space of $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$, but for $N_{tb} > K$ more than one vector lie in the null space of $\left\{\boldsymbol{h}_{b,j,l}^H\right\}_{j\neq k}$. In this latter case, the final $\boldsymbol{w}_{b,k,l}$ vector is a linear combination of the $\left(N_{t_b} - K + 1\right)$ possible solutions. The equivalent channel between BS b and UT k, on sub-carrier l can be expressed as,

$$\boldsymbol{h}_{b,k,l}^H \boldsymbol{w}_{b,k,l} = \boldsymbol{h}_{b,k,l}^H \bar{\mathbf{W}}_{b,k,l} \frac{\left(\boldsymbol{h}_{b,k,l}^H \bar{\mathbf{W}}_{b,k,l}\right)^H}{\left\|\boldsymbol{h}_{b,k,l}^H \bar{\mathbf{W}}_{b,k,l}\right\|} = \left\|\boldsymbol{h}_{b,k,l}^H \bar{\mathbf{W}}_{b,k,l}\right\| = h_{b,k,l}^{eq} \tag{26}$$

From (26) we can observe that the equivalent channel, $h_{b,k,l}^{eq}$, is a positive real number. By using the precoding vectors defined in (25) and considering (26), the received signal in (24) reduces to,

$$y_{k,l} = \sum_{b=1}^{B} \sqrt{p_{b,k,l}} h_{b,k,l}^{eq} s_{k,l} + n_{k,l} \tag{27}$$

It should be mentioned that at the UT, to allow high order modulations, only the $\sqrt{p_{b,k,l}}h_{b,k,l}^{eq}$ coefficients are needed to be estimated instead of all the complex coefficients of the channel, leading to a low complexity UT design.

Since the $\left(N_{t_b}-K+1\right)$ components of $\boldsymbol{h}_{b,k,l}^{H}\overline{\mathbf{W}}_{b,k,l}$ are i.i.d. Gaussian variables, $\left(h_{b,k,l}^{eq}\right)^2$ is a chi-square random variable with $2\left(N_{t_b}-K+1\right)$ degrees of freedom. Once the $h_{b,k,l}^{eq}$ variables are independent, each user is expected to achieve a diversity order of $B\left(N_{t_b}-K+1\right)$ (assuming that all channels have the same average power, i.e., $\rho_{b,k}=\rho,\ \forall(b,k)$ and $p_{b,k,l}=1,\ \forall(b,k,l)$). Also, because the received signals from different BSs have the same phase, they are added coherently at the UTs, and thus an additional antenna gain is achieved.

4.3 Power allocation strategies

In this section the same three criteria considered for the centralized approach are used to design the power allocation. However, it should be emphasised that for this scenario only the equivalent channels, i.e., $h_{b,k,l}^{eq}$, are needed to be known at the JPU.

A. Optimal minimum BER power allocation

From (27) the instantaneous SNR of user k on sub-carrier l can be written as,

$$\mathrm{SNR}_{k,l}=\frac{\left(\sum_{b=1}^{B}\sqrt{p_{b,k,l}}h_{b,k,l}^{eq}\right)^2}{\sigma^2} \tag{28}$$

The instantaneous probability of error for user k is obtained in similar way in section 3. We minimize the instantaneous average probability under the per-BS power constraint P_{t_b}, i.e., $\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}\le P_{t_b},\ b=1,\dots,B$. By assuming a 4-QAM constellation, the optimal power allocation problem with per-BS power constraint can be formulated as,

$$\min_{\{p_{b,k,l}\}}\left(\frac{1}{KN_c}\sum_{l=1}^{N_c}\sum_{k=1}^{K}Q\left(\frac{\sum_{b=1}^{B}\sqrt{p_{b,k,l}}h_{b,k,l}^{eq}}{\sigma}\right)\right)\ \text{s.t.}\ \begin{cases}\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}\le P_{t_b},b=1,..,B\\ p_{b,k,l}\ge 0,\ \ b=1,..,B,k=1,..,K,l=1,..,N_c\end{cases} \tag{29}$$

In this distributed approach, the objective function is convex in $p_{b,k,l}$, and the constraint functions are linear this is also a convex optimization problem. Therefore, it may be also solved numerically by using for example the interior-point method. This scheme is referred as distributed per-BS optimal power allocation (Dist. per-BS DOPA). In this section, the distributed term is referred to the precoder vectors since the power allocation is also computed in a centralized manner.

B. Suboptimal power allocation approaches

As for the centralized approach, the complexity of the above scheme is too high, and thus it is not of interest for real wireless systems, we also resort to less complex suboptimal solutions. The proposed strategy has two phases: first the power allocation is computed by assuming that all BSs of each super-cell can jointly pool their power, i.e., a TPC P_t is imposed instead and the above optimization problem reduces to,

$$\min_{\{p_{b,k,l}\}}\left(\frac{1}{KN_c}\sum_{l=1}^{N_c}\sum_{k=1}^{K}Q\left(\frac{\sum_{b=1}^{B}\sqrt{p_{b,k,l}}h_{b,k,l}^{eq}}{\sigma}\right)\right)\text{ s.t. }\begin{cases}\sum_{b=1}^{B}\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}\le P_t\\ p_{b,k,l}\ge 0,\ b=1,..,B,k=1,..,K,l=1,..,N_c\end{cases}\tag{30}$$

with $P_t=\sum_{b=1}^{B}P_{t_b}$ and using the Lagrange multipliers method, the following cost function with μ Lagrange multiplier is minimized,

$$J_{d,1}=\frac{1}{KN_c}\sum_{l=1}^{N_c}\sum_{k=1}^{K}Q\left(\frac{\sum_{b=1}^{B}\sqrt{p_{b,k,l}}h_{b,k,l}^{eq}}{\sigma}\right)+\mu\left(\sum_{b=1}^{B}\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}-P_t\right)\tag{31}$$

The powers $p_{b,k,l}$, $\forall(b,k,l)$ can be determined by setting the partial derivatives of $J_{d,1}$ to zero and as shown in (Silva et al., 2011) the solution is,

$$p_{b,k,l}=\frac{\sigma^2\left(h_{b,k,l}^{eq}\right)^2}{\left(\sum_{i=1}^{B}\left(h_{i,k,l}^{eq}\right)^2\right)^2}W_0\left(\frac{\left(\sum_{i=1}^{B}\left(h_{i,k,l}^{eq}\right)^2\right)^2}{8\pi\mu^2N_c^2K^2\sigma^4}\right)\tag{32}$$

Therefore, μ^2 can be determined iteratively, using constraint $\sum_{b=1}^{B}\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}=P_t$. The second phase consists of replacing μ^2 by μ_b^2, $b=1,...,B$ in (32), and then computing iteratively different μ_b^2 to satisfy the individual per-BS power constraints instead, i.e., μ_b^2 are computed to satisfy,

$$\begin{cases}\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}\le P_{t_b},\ b=1,...,B\\ p_{b,k,l}\ge 0,\ b=1,..,B,k=1,..,K,l=1,..,N_c\end{cases}\tag{33}$$

This suboptimal scheme is referred as distributed per-BS sub-optimal iterative power allocation (Dist. per-BS SOIPA). Although this suboptimal solution significantly reduces the complexity relative to the optimal one, it still needs an iterative search. To further simplify we also propose for the distributed scenario, an alternative power allocation method based on minimizing the sum of inverse of SNRs.

In this case, the optimization problem is written as,

$$\min_{\{p_{b,k,l}\}}\left(\sum_{l=1}^{N_c}\sum_{k=1}^{K}\frac{\sigma^2}{\left(\sum_{b=1}^{B}\sqrt{p_{b,k,l}}h_{b,k,l}^{eq}\right)^2}\right)\text{ s.t. }\begin{cases}\sum_{l=1}^{N_c}\sum_{k=1}^{K}p_{b,k,l}\le P_{t_b},b=1,..,B\\ p_{b,k,l}\ge 0,\ b=1,..,B,k=1,..,K,l=1,..,N_c\end{cases}\tag{34}$$

The objective function is convex in $p_{b,k,l}$, and the constraint functions are linear, (34) is also a convex optimization problem. To solve it we follow the same suboptimal two phases approach as for the first problem.

First, we impose a total power constraint and the following cost function, using again the Lagrangian multipliers method, is minimized,

$$J_{d,2} = \sum_{l=1}^{N_c}\sum_{k=1}^{K} \frac{\sigma^2}{\left(\sum_{b=1}^{B}\sqrt{p_{b,k,l}}\,h_{b,k,l}^{eq}\right)^2} + \mu\left(\sum_{b=1}^{B}\sum_{l=1}^{N_c}\sum_{k=1}^{K} p_{b,k,l} - P_t\right) \tag{35}$$

Now, setting the partial derivatives of $J_{d,2}$ to zero and after some mathematical manipulations, the powers $p_{b,k,l}$ can be shown to be given by,

$$p_{b,k,l} = \frac{\left(h_{b,k,l}^{eq}\right)^2}{\beta\sqrt{\left(\sum_{i=1}^{B}\left(h_{i,k,l}^{eq}\right)^2\right)^3}} \tag{36}$$

where $\beta = \sqrt{\mu / \sigma^2}$. As for the first approach, (36) can be re-written by replacing β by β_b, $b = 1,...,B$, which are computed to satisfy the individual per-BS power constraints and the closed-form solution achieved is then given by,

$$p_{b,k,l} = \frac{P_{t_b}\left(h_{b,k,l}^{eq}\right)^2}{\sqrt{\left(\sum_{i=1}^{B}\left(h_{i,k,l}^{eq}\right)^2\right)^3}\sum_{p=1}^{N_c}\sum_{j=1}^{K}\frac{\left(h_{b,j,p}^{eq}\right)^2}{\sqrt{\left(\sum_{i=1}^{B}\left(h_{i,j,p}^{eq}\right)^2\right)^3}}} \tag{37}$$

This second suboptimal scheme is referred as distributed per-BS closed-form power allocation (Dist. per-BS SOCPA).

The precoder vectors are designed by assuming that BSs have only knowledge of local CSI. However, since we consider a centralized power allocation, to compute all powers the $h_{b,k,l}^{eq}$, $\forall b,k,l$ coefficients should be available at the joint processing unit (JPU). In the distributed multi-cell system each BS should send a real vector of size KN_c to the JPU. Note that in the centralized approach discussed in section 3, each BS should send to the JPU a complex vector of size $N_{t_b}KN_c$, i.e. $2N_{t_b}$ more information.

Although, in this section single antenna UTs were assumed, the formulation can be straightforwardly extended for multiple antenna UTs just by considering each antenna as a single antenna UT. The main difference is that the long term channel power will be the same for all antennas belonging to the same UT.

5. Results and discussions

5.1 Simulation parameters

In order to evaluate the proposed centralized and distributed multi-cell cooperation schemes, we assume ITU pedestrian channel model B (Guidelines IMT2000, 1997), with the modified taps' delays, used according to the sampling frequency defined on LTE standard (3GPP LTE, 2007). This time channel model was extended to space-time by assuming that the distance between antenna elements of each BS is far apart to assume uncorrelated channels. To evaluate centralized and distributed schemes, the follwoing scenarios are considered:

- Scenario 1, we assume that each supercell has 2 BSs, $B = 2$ which are equipped with 2 antennas, $N_{t_b} = 2$ and 2 UTs, $K = 2$, equipped with 2 antennas, $N_{r_k} = 2$.
- Scenario 2, we assume that each supercell has 2 BSs, $B = 2$ which are equipped with 2 antennas, $N_{t_b} = 2$ and 2 single antenna UTs, $K = 2$.
- Scenario 3, we assume that each supercell has 2 BSs, $B = 2$ which are equipped with 4 antennas, $N_{t_b} = 4$ and 2 single antenna UTs, $K = 2$.

The main parameters used in the simulations are, FFT size of 1024; number of resources, i.e., available subcarriers (N_c) shared by the K users set to 16; sampling frequency set to 15.36 MHz; useful symbol duration is 66.6 μs; cyclic prefix duration is 5.21 μs; overall OFDM symbol duration is 71.86 μs; subcarrier separation is 15 kHz and modulation is 4-QAM. We assume that each UT is placed on each cell. The long-term channel powers are assumed to be $\rho_{b,k} = 1,\ b = k$ for the intracell links, and $\rho_{b,k},\ b \neq k$ are uniformly distributed on the interval $[0.2\ ,\ 0.6]$ for the intercell links. All the results are presented in terms of the average BER as a function of per-BS SNR defined as $SNR = P_{tb} / \sigma^2$.

5.2 Performance evaluation

5.2.1 Centralized scenario

This section presents the performance results of centralized proposed precoding approaches for scenario 1. We compare the performance results of four centralized precoding schemes: one with non power allocation, which is obtained for the single cell systems by setting $\mathbf{P}_l = \mathbf{I}_{N_r}$, i.e., the power per data symbol is constrained to one. For multi-cell systems the power matrix $\mathbf{P}_l = \mathbf{I}_{N_r}$ should be scaled by β as defined in (16) (setting $p_{k,i,l} = 1, \forall k,i,l$), i.e., $\mathbf{P}_l = \beta \mathbf{I}_{N_r}$ ensuring a per-BS power constraint instead. This scheme is referred as centralized per-BS non-power allocation (Cent. per-BS NPA). The two suboptimal approaches are Cent. per-BS SOCPA and Cent. per-BS SOIPA; and the optimal one is Cent. per-BS OPA. Also, we present results for optimal approach considering total power allocation (Cent. TPC OPA), as formulated in (13), which may serve as a lower bound of the average BER for the centralized multi-cell system with per-BS power constraint.

Fig. 3 shows the performance results of all considered precoding schemes for scenario 1, considering multiplexing mode. It can be observed that the Cent. per-BS SOCPA, Cent. per-BS SOIPA and Cent. per-BS OPA schemes have significant outperformance comparing to the Cent. per-BS NPA approach, because they redistribute the powers across the different subchannels more efficiently. Comparing the two suboptimal approaches we can see that the iterative one, Cent. per-BS SOIPA, outperforms the closed-form, Cent. per-BS SOCPA because the former is obtained by explicitly minimizing average probability of error. The performance of the proposed suboptimal Cent. per-BS SOIPA and Cent. per-BS SOCPA approaches is close, a penalty less than 0.7 dB for a BER=10^{-2} can be observed. Also, the penalty of the Cent. per-BS SOIPA against the lower bound given by the Cent. TPC OPA is only about 0.5 dB considering also a target BER=10^{-2}.

Fig. 4 shows the performance results of all considered precoding schemes for scenario 1, considering diversity mode. Comparing these results with the last ones, it can be easily seen that there is a large gain due to operating in diversity mode. Since now each data symbol is

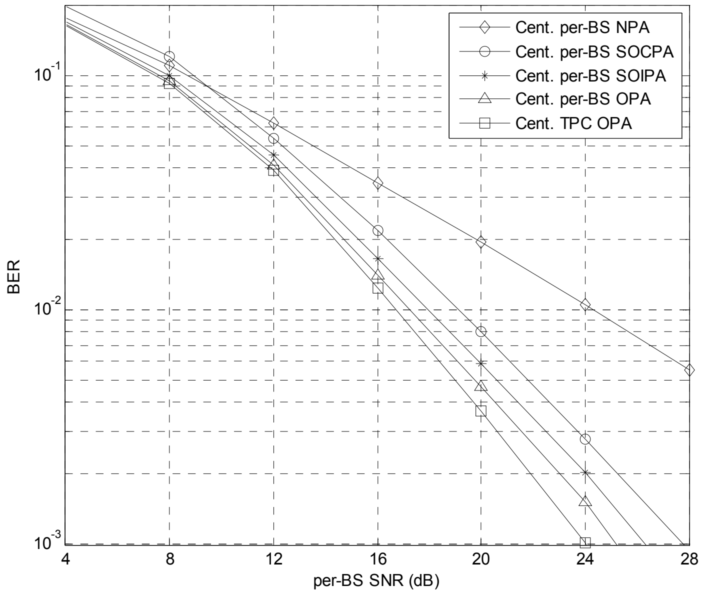


Fig. 3. Performance evaluation of the proposed centralized multi-cell schemes considering multiplexing mode, for scenario 1

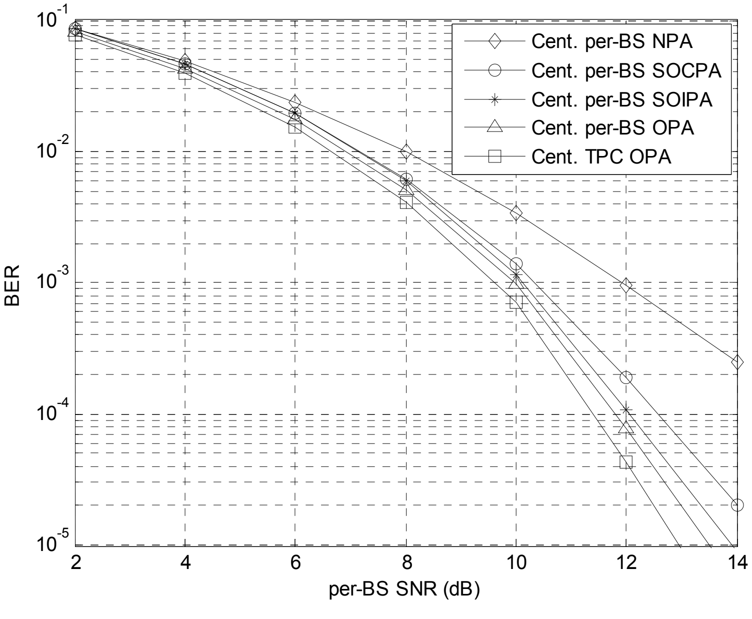


Fig. 4. Performance evaluation of the proposed centralized multi-cell schemes considering diversity mode, for scenario 1

collected by each receive antenna of each UT. From this figure we basically can point out the same conclusions as for the results obtained in the previous one. However, one important thing that can be found out by comparing multiplexing and diversity modes is that the difference between Cent. per-BS NPA curves and power allocation based curves (e.g. Cent. per-BS SOIPA) is bigger in multiplexing mode (approximately 4dB) than diversity mode (1.5dB) considering a BER=10^{-2}. This can be explained by the fact that in the diversity mode the equivalent channel gain of each data symbol is the addition of N_{r_k} individual channel gains and thus the dynamic range of the SNRs of the different data symbols is reduced, i.e., somewhat leads to an equalization of the SNRs.

5.2.2 Distributed scenario

This section presents the performance results of proposed distributed precoding approaches for scenario 2. We compare the results of four distributed precoding schemes with different per-BS power allocation approaches: distributed per-BS equal power allocation (Dist. per-BS EPA), in this case $p_{b,k,l} = P_{t_b} / KN_c ,\forall(b,k,l)$; the two suboptimal approaches Dist. per-BS SOIPA and Dist. per-BS SOCPA and the optimal one Dist. per-BS OPA. Also, the results for optimal approach considering total power allocation (Dist. TPC OPA) , as formulated in (30) are presented. This serves as lower bound for the distributed multi-cell scenario under per-BS power constraint.

Fig. 5 shows the performance results of all considered distributed precoding schemes for scenario 2. It can be observed that the Dist. per-BS SOCPA, Dist. per-BS SOIPA and Dist. per-BS OPA schemes outperform the Dist. per-BS EPA approach, because they redistribute the powers across the different subchannels more efficiently. For this case the performance

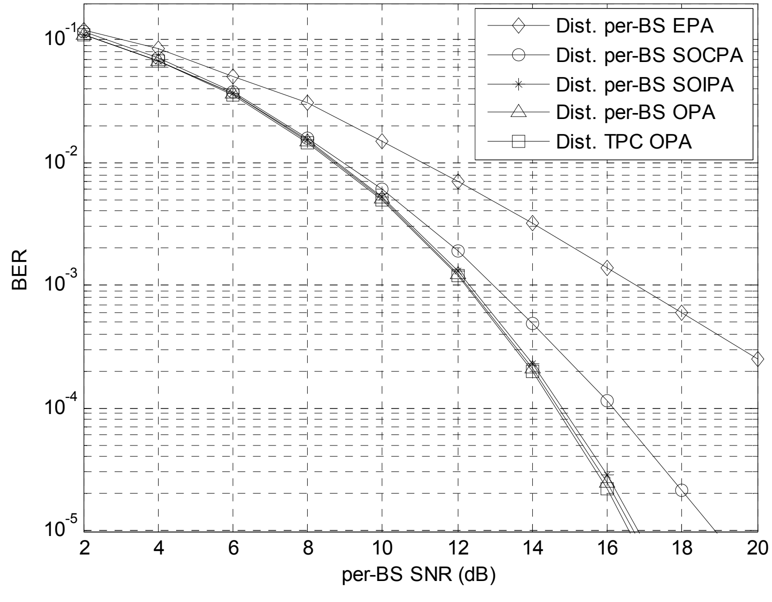


Fig. 5. Performance evaluation of the proposed distributed multi-cell schemes, for scenario 2

of the suboptimal Dist. per-BS SOIPA and optimal Dist. per-BS OPA is very close (penalty less than 0.1dB), but the gap between these two schemes and the suboptimal Dist. per-BS SOCPA is considerable. These results show that the Dist. per-BS SOIPA outperforms the Dist. per-BS SOCPA for large number of subchannels. We can observe a penalty of approximately 0.6 dB of the Dist. per-BS SOCPA scheme against the Dist. per-BS SOIPA for a BER=10^{-3}. Also, a gain of approximately 4.2 dB of the suboptimal Dist. per-BS SOIPA scheme against the Dist. per-BS EPA is obtained, considering BER=10^{-3}.

5.2.3 Performance comparison

This section presents the performance results of both distributed and centralized proposed precoding approaches for scenarios 2 and 3.

Fig. 6 shows the results for scenario 2, from this figure we can see that the performance of all power allocation schemes with centralized precoding outperforms the one with distributed scheme, because there are more degrees of freedom (DoF) to remove the interference and enhance the system performance. In the distributed case, the performance of the suboptimal Dist. per-BS SOIPA and optimal Dist. per-BS OPA is very close (penalty less than 0.1dB), but the gap between these two schemes and the suboptimal per-BS SOCPA is almost increased to 0.8dB (BER=10^{-3}). In the case of centralized precoding the performances of Cent. per-BS SOIPA and Cent. per-BS OPA are still very close but both are degraded from Cent. TPC OPA (about 0.5dB at BER=10^{-3}) and also there is 0.5dB gap among these curves and Cent. per-BS SOCPA at the same BER. Another important issue that should be emphasized is that the penalty of the per-BS OPA against the TPC OPA is approximately 0.1 dB (BER=10^{-3}) for distributed scheme, against 0.5dB for centralized case.

Figure 7 shows the performance results of both distributed and centralized schemes for scenario 3. By observing this figure almost the same conclusions can be drawn. An interesting result is that the performances of distributed and centralized schemes are much closer comparing with scenario 2. This can be explained by the fact that for the centralized approach the number of DoF, which is given by the number of total transmit antennas BN_{t_b}, increased from 4 (scenario 2) to 8 (scenario 3); while for the distributed approach, the number of DoF, which is given by $B\left(N_{t_b} - K + 1\right)$ as discussed before; is increased from 2 (scenario 2) to 6 (scenario 3), i.e., the number of DoF of both centralized and distributed approaches is closer than that in scenario 2. From the presented results two important facts should be also emphasized: first is that in case of distributed precoding, the performance improvement achieved with the three proposed power allocation techniques, is higher than the case of centralized scheme; the second is that in the case of distributed precoding, the suboptimal techniques are more successful in achieving the lower bound of average BER.

6. Conclusion

In this chapter we proposed and evaluated centralized and distributed multi-cell multiuser precoding schemes for MIMO OFDM based systems. The proposed precoder vectors were computed either jointly and centraly at JPU benefiting from high DoF or on each BS in a distributed manner allowing a low feedback load over the backhaul network, while the power allocation was computed in a centralized fashion at the JPU.

The criteria considered was the minimization of the BER and two centralized power allocation algorithms with per-BS power constraint: one optimal that can be achieved at the expense of some complexity and one suboptimal with lower complexity aiming at practical

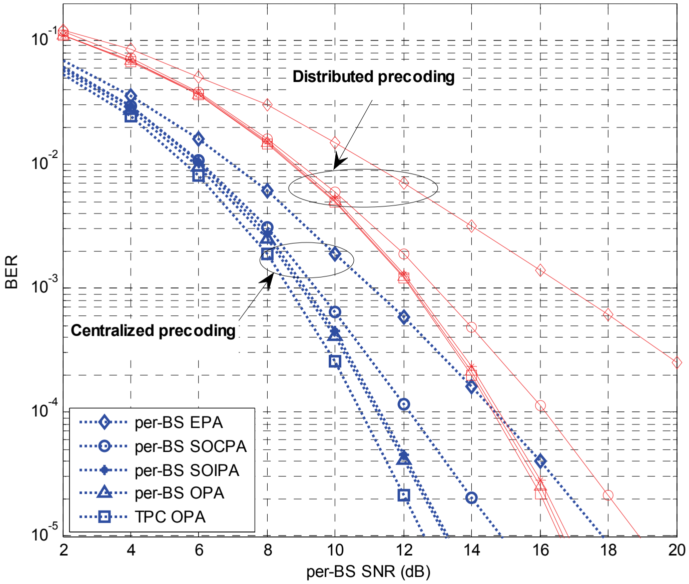


Fig. 6. Performance evaluation of the proposed distributed and centralized multi-cell schemes for scenario 2

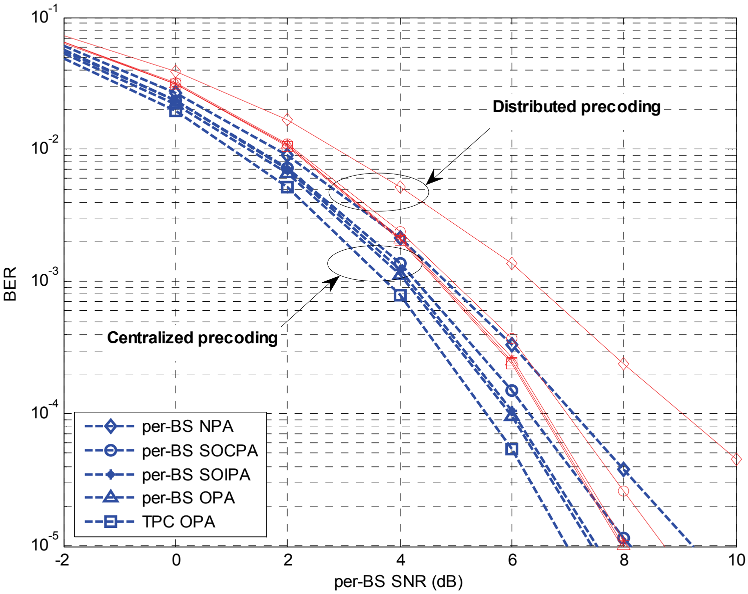


Fig. 7. Performance evaluation of the proposed distributed and centralized multi-cell schemes for scenario 3

implementations. In both the optimal (per-BS OPA) and the suboptimal (per-BS SOIPA), the computation of the transmitted powers required an iterative approach. To circumvent the need for iterations further proposed another suboptimal scheme (per-BS SOCPA), where the power allocation was computed in order to minimize the sum of inverse of SNRs of each UT allowing us to achieve a closed-form solution.

The results have shown that the proposed multi-user multi-cell schemes cause significant improvement in system performance, in comparison with the case where no power allocation is used. Also for both approaches, the performance of the proposed suboptimal algorithms, namely the per-BS SOIPA approach, is very close to the optimal with the advantage of lower complexity. Also, the performance of the distributed approach tends to the one achieved by the centralized, when the number of DoF available tends to the number of DoF available in the centralized system. Therefore, distributed schemes can be interesting in practice when the backhaul capacity is limited.

It is clear from the presented results the suboptimal proposed either distributed or centralized precoding schemes allow a significant performance improvement with very low UT complexity and moderate complexity at both BS and JPU, and therefore present significant interest for application in next generation wireless networks for which cooperation between BSs is anticipated.

7. Acknowledgments

The authors wish to acknowledge the support of the Portuguese CADWIN project, PTDC/EEA TEL/099241/2008, and Portuguese Foundation for Science and Technology (FCT) grant for the second author.

8. References

Andrews, J. G.; Choi, W. & Heath, R. W. (2007). Overcoming interference in spatial multiplexing MIMO cellular networks, *IEEE Wireless Communication Magazine*, vol. 47, no. 6, pp. 95-104, 2007.

Armada, A. G.; Fernándes, M. S. & Corvaja, R. (2009). Waterfilling schemes for zero-forcing coordinated base station transmissions, in *Proceeding of IEEE GLOBECOM*, 2009.

Bahai, A. R. S.; Saltzberg, B. R. & Ergen, M. (2004). *Multi-carrier digital communications theory and applications of OFDM*, Springer, New York.

Bjornson, E.; Zakhour, R.; Gesbert, D. & Ottersten, B. (2010). Cooperative multi-cell precoding: rate region characterization and distributed strategies with instantaneous and statistical CSI, *IEEE Transaction on Signal Processing*, vol. 58, no. 8, pp. 4298-4310, 2010.

Boccardi, F. & Huang, H. (2007). Limited downlink network coordination in cellular networks, *In Proeeding IEEE PIMRC'07*, 2007.

Boyd, S. & Vandenberghe, L. (2004). *Convex optimization*, Cambridge: Cambridge University Press, 2004.

Corless, R. M.; Gonnet, G. H.; Hare, D. E. G.; Jeffrey, D. J. & Knuth, D. E. (1996). On the Lambert W function, *Advances in Computer & Mathematics,* vol. 5, pp. 329–359, 1996.

Costa, M. H. M. (1983). Writing on dirty paper, *IEEE Transaction on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.

Diehm, F.; Marsch, P. & Fettweis, G. (2010). The FUTON prototype: proof of concept for coordinated multi-point in conjunction with a novel integrated wireless/optical architecture, *In Proceeding of IEEE WCNC'10*, 2010.

Foschini, G. J. & Gans, M. J. (1998). On limits of wireless communications in a fading environment when using multiple antenna, *Wireless Personal Communication Magazine*, vol. 6, no. 3, pp. 311-335, 1998.

Foschini, G. J. et al. (2006). Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency, *in IEEE Proceedings on Communications*, Vol. 153, No. 4, pp. 548-555, August 2006.

FUTON, European ICT project. (2007). Available from: https://www.ict-futon.eu.

Gesbert, D.; Hanly, S.; Huang, H.; Shamai, S.; Simeone, O. & Yu, Wei (2010). Multi-cell MIMO cooperation networks: A new look at interference, *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 9, pp. 1380-1408, Dec. 2010.

Guidelines for the evaluation of radio transmission technologies for IMT-2000. (1997). Recommendation ITU-R M.1225, 1997.

Haykin, S. (1996). *Adaptive Filter Theory*, 3rd Ed., Prentice Hall, 1996.

Holakouei, R; Silva, A & Gameiro, A. (2011). Multiuser precoding techniques for a distributed broadband wireless system, *Accepted to Telecommunication System Journal, special issue in WMCNT, Springer*.

Silva, A; Holakouei, R & Gameiro, A. (2011). Power Allocation Strategies for Distributed Precoded Multicell Based Systems, *EURASIP Journal on Wireless Communications and Networking, special issue in Recent Advances in Multiuser MIMO System,* Vol. 2011, 2011.

Jing, S.; Tse, D. N. C.; Sorianga, J. B.; Hou, J.; Smee, J. E. & Padovani, R. (2008). Multicell downlink capacity with coordinated processing, *EURASIP Journal of Wireless Communicaion Network*, 2008.

Karakayali, M.; Foschini, G. & Valenzuela, G. (2006). Network coordination for spectral efficient communications in cellular systems, *IEEE Wireless Communication Magazine*, vol. 13, no. 4, pp. 56-61, 2006.

Kobayashi, M.; Debbah, M. & Belfiore, J. (2009). Outage efficient strategies in network MIMO with partial CSIT, in *Proceeding of IEEE ISIT,* pp. 249-253, 2009.

Lozano, A.; Tulino, A. M. & Verdú, S. (2008). Optimum Power Allocation for Multiuser OFDM with Arbitrary Signal Constellations, *IEEE Transactions on Communications*, vol. 56, no. 5, pp. 828-837, 2008.

Marsch, P. & Fettweis, G. (2009). On Downlink network MIMO under a constrained backhaul and imperfect channel knowledge, in *Proceeding of IEEE GLOBECOM*, 2009.

Mudumbai, T. R.; Brown, D. R.; Madhow, U. & Poor, H. V. (2009). Distributed transmit beamforming: challenges and recent progress, *IEEE Communication Magazine*, vol. 47, no. 2, pp. 102-110, 2009.

Palomar, D. P.; Cioffi, J. M. & Lagunas, M. A. (2003). Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization, *IEEE Transactions on Signal Processing,* vol. 51, no. 9, pp. 2381-2401, 2003.

Proakis, J. (1995). *Digital Communications*, 3 rd Ed., McGrraw-Hill, New York, 1995.

Rostaing, P.; Berder, O.; Burel, G. & Collin, L. (2002). Minimum BER diagonal precoder for MIMO digital transmissions, *Elsevier Signal Processing*, vol. 82, no. 10, pp. 1477-1480, 2002.

Somekh, O.; Zaidel, B. & Shamai, S. (2007). Sum rate characterization of joint multiple cell-site processing, *IEEE Transaction of Information Theory*, vol. 53, no. 12, pp. 4473- 4497, 2007.

Third Generation Partnership Project, (2007). LTE Physical Layer - General Description, no 3. 3GPP TS 36.201 v8.1.0, 2007. Available from: http://www.3gpp.org/LTE.

Uppala, L. S. & Li, J. (2004). Designing a mobile broadband wireless access network, *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 20-28, 2004.

Zhang, J.; Chen, R.; Andrews, J. G.; Ghosh, A. & Heath Jr., R. W. (2009) Networked MIMO with Clustered Linear Precoding, *IEEE Transaction on Wireless Communication*, vol. 8, no. 4, pp. 1910-1921, 2009.

Zhang, R. (2010). Cooperative multi-cell block diagonalization with per-base-station power constraints, in *Proceeding of IEEE WCNC*, 2010.

Part 2

Upper Layers

9

Joint Call Admission Control in Integrated Wireless LAN and 3G Cellular Networks

Chunming Liu, Chi Zhou, Niki Pissinou and S. Kami Makki
T-Mobile, Illinois Institute of Technology, Florida International University, Lamar University
U.S.A.

1. Introduction

The fourth-generation (4G) (Liu, 2004) system is expected to support fully integrated services and ubiquitous access anytime and anywhere. Instead of developing a new uniform standard for all wireless communication systems, some endeavors in 4G research focus on the seamless integration of various existing wireless communication networks, such as integrated Wireless LAN (WLAN) and the third-generation (3G) cellular networks.

3G cellular networks provide wide coverage and universal roaming services with limited data rate up to 2 Mbps (Liu, 2006, 2007). With careful network planning and mature admission control algorithms, the achievable Quality of Service (QoS) level of 3G cellular networks is relatively high. On the other hand, WLANs provide low-cost, high data rate wireless access within limited hotspot-area. Since WLAN is originally designed for best-effort data services with contention-based access, it is difficult to achieve strict QoS provisioning for real-time services, such as voice service (Song et al., 2006).

Due to different network capacities, user mobile patterns, vertical handoffs, and QoS levels, the integrated WLAN and 3G cellular networks require a new call admission control scheme to provide QoS provisioning and efficient resource utilization. Currently there are three major architectures for internetworking between 3G cellular cellular networks and WLAN (Ahmavaara et al., 2003). But they are all lack of joint resource management and admission control schemes in integrated environment. Previous research work on admission control in homogeneous cellular networks and heterogeneous integrated networks are investigated with technical descriptions on their pros and cons. It is shown that more endeavors are needed on joint congestion control, load balance, and high-level QoS provisioning in integrated networks.

In this chapter, a novel joint call admission control (CAC) scheme is proposed to support both voice and data services with QoS provisioning. Due to different network service characteristics, 3G cellular network is defined to be a voice-priority network where voice services have higher priority for resource allocation than data services, while WLAN is defined as data-priority network where data services have higher priority than voice services. A joint call admission policy is derived to support heterogeneous network architecture, service types, QoS levels, and user mobility characteristics. Furthermore, to relieve traffic congestion in cellular networks, an optimal channel searching and replacement algorithm and related passive handoff techniques are further developed to balance total system traffic between WLAN and 3G cellular network, as well as to reduce average system QoS cost, such as system blocking probability. A one-dimensional Markov model for voice service is also developed to analyze interworking system performance metrics. Both theoretical analysis and simulation results show that average system QoS costs, such as overall blocking and dropping probabilities, are reduced, and our scheme outperforms both traditional disjoint static CAC scheme and joint CAC without optimization.

2. Technical background

This section briefly describes concepts, architecture and vertical handoffs in integrated WLAN and cellular networks.

2.1 Architecture of integrated WLAN and 3G cellular networks

Driven by the anywhere and anytime mobile service concept, it is expected that 4G wireless networks will be heterogeneous, integrating different networks to provide seamless Internet access for mobile users. The integrated WLAN and 3G cellular network takes advantage of the wide coverage and almost universal roaming support of 3G cellular networks and the high data rates of WLANs.

Currently, there are three major architectures for internetworking between 3G Universal Mobile Telecommunications System (UMTS) cellular networks and IEEE 802.11 WLAN. These are Open Coupling, Tight Coupling, and Loose Coupling (Liu, 2006). The Open Coupling architecture specifies an open standard and is used for access and roaming between 802.11 WLAN and UMTS networks. In this approach, both networks are considered as two independent systems that may share a single billing scheme between them. An 802.11 WLAN is connected to the Internet through a Gateway Router, and UMTS network, is connected to the Internet through a Gateway GPRS Support Node (GGSN). Open Coupling scheme is lack of supports for mobility, resource management, QoS provisioning, and security in integrated environment.

As a direct integration scheme, Tight Coupling connects the WLAN network to the rest of the core network in the same manner as other cellular radio access technologies (Liu & Zhou, 2005a, 2005b; Liu, 2006). As shown in Fig. 1, the WLAN gateway router hides the details of the WLAN from the 3G UMTS core network by adding a new component, SGSN emulator, into WLAN. The SGSN emulator connects the gateway router in the WLAN to the IP core network. It interconnects the UMTS core network at the $\mathbf{G_n}$ interface (Liu, 2006), and implements all UMTS protocols required in a 3G radio access network. In terms of UMTS protocols, the WLAN service area works like another SGSN coverage area to the UMTS core network. As a result, all the traffics, including data and UMTS signaling, generated in the WLAN are injected directly into the UMTS core network through the SGSN emulator. This increases the traffic load of the UMTS core network. If the operators of the WLAN are different from those of UMTS network, the new interface between the UMTS and the WLAN can cause security weaknesses. In addition, the WLAN cards in client devices must incorporate the UMTS protocol stack, and Universal Subscriber Identity Module (USIM) authentication mechanism must be used for authentication in the WLAN (Liu & Zhou, 2005a).

In contrast to high cost of Tight Coupling, the Loose Coupling is an IP-based mechanism, and approach separates the data paths in the 802.11 WLAN and 3G cellular networks (Liu, 2006). The 802.11 WLAN gateway routers connect to the Internet, and all data traffic is

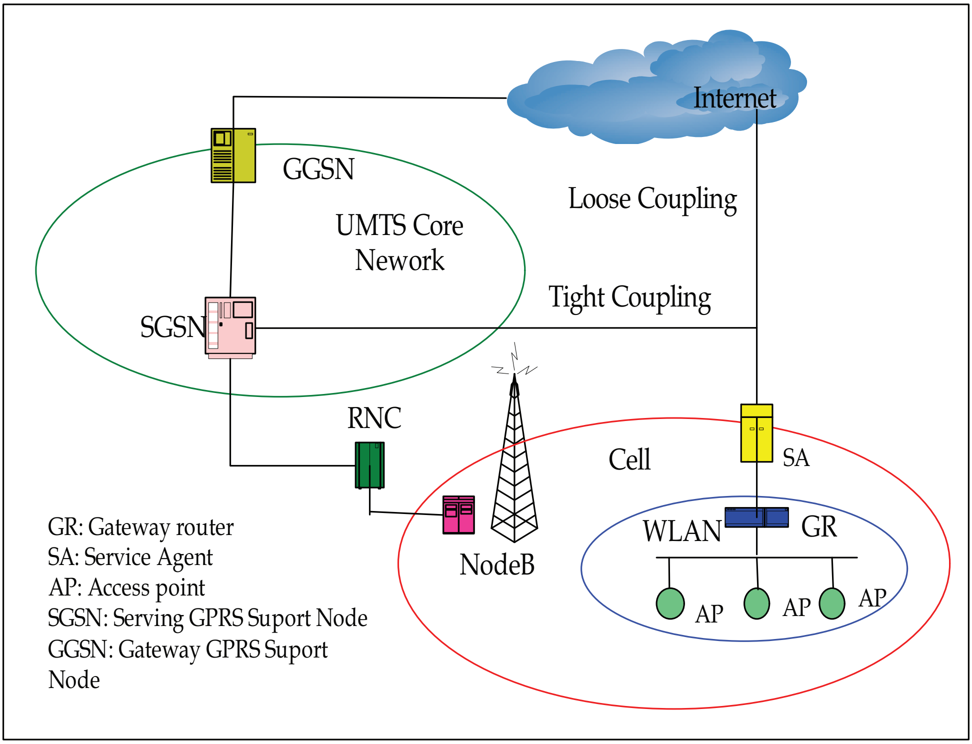


Fig. 1. Tight coupling and loose coupling

routed to the core Internet, instead of to the cellular core network. To the core network of the UMTS for example, the 802.11 WLAN appears like a visiting network. The gateway of the 802.11 WLAN can be connected to a Service Agent (SA), a combined SGSN/GGSN emulator, which provides not only internetworking protocol for signaling between the 802.11 WLAN and the UMTS 3G core network, but also an interface for data traffics between the WLAN and IP networks. If the 802.11 WLAN is deployed by the same UMTS operator, the SA may interface directly to the UMTS Core Network for signaling. Otherwise, the SA is interfaced to the IP network for both signaling and data traffic. Compared to open or tight coupling architectures, loose coupling implements the independent deployment and traffic configuration of both the 802.11 WLAN and UMTS networks. In addition, loose coupling architecture allows a mobile operator to provide its own private 802.11 WLAN “hotspots” and interoperate with public 802.11 WLANs and UMTS operators via internetworking agreements. So generally speaking, loose coupling is most preferable for integrated WLAN/Cellular network, due to the simplicity and less reconfiguration work.

Though promising, loose coupling have several technical open issues to be addressed before successful integration, such as integrated location management, seamless vertical handoff, common QoS provisioning, unified Authentication, Authorization and Accounting (AAA), joint call admission control and so on. As a part of resouce management, joint call admission control tightly interacts with vertical handoff and QoS provisioning schemes in integrated WLAN and 3G cellular networks.

2.2 Vertical handoff

In integrated networks, there are two types of handoff: intra-technology handoff and inter-technology handoff (Lampropoulos et al., 2005; Shafiee et al., 2011). The intra-technology handoff is traditional Horizontal Handoff (HHO) in which mobile terminals handoff between two adjacent base stations or access points using same access technology. In contrast, inter-technology handoff is called Vertical Handoff (VHO), and happens when mobile terminals roam between two networks with different access technologies, for example, between WLAN and 3G UMTS network.

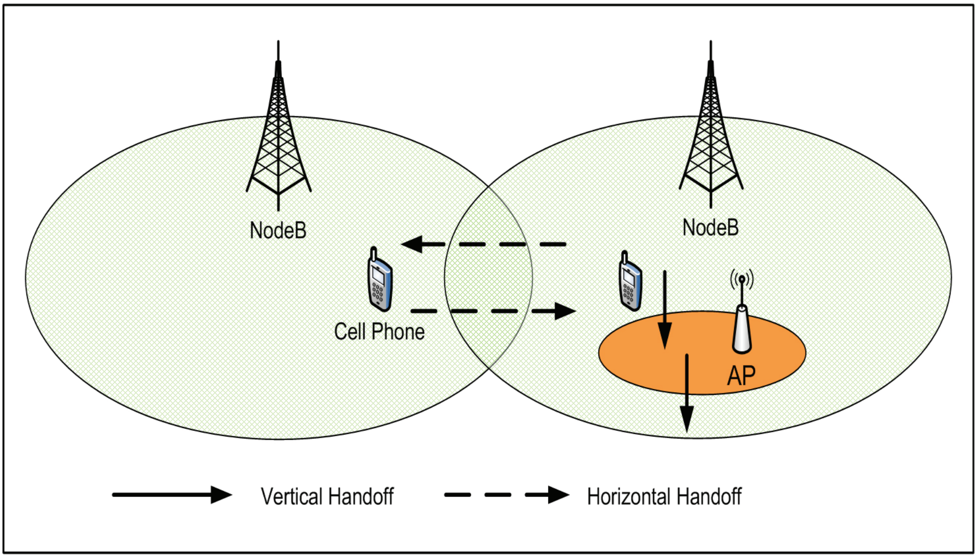


Fig. 2. Handoffs in integrated WLAN and UMTS cellular networks

Vertical handoffs in integrated WLAN / UMTS networks have two scenarios: a mobile terminal moves out of a WLAN to a UMTS cellular network, and moves from UMTS cellular network into a WLAN. Considering different service coverage area, the vertical handoff from WLAN to Cellular network is normally triggered by signal fading when a user moves out of the service area of the WLAN. However, the vertical handoff from cellular network to WLAN is regarded as a network selection process, because mobile terminals are in a wireless overlay area where both cellular access and WLAN access are available to mobile terminals at same time.

Seamless vertical handoffs face challenges caused by the gap between different QoS levels in cellular network and in WLAN (Liu, 2006; Shafiee et al., 2011): UMTS cellular networks provide wide coverage with high QoS provisioning for voice service, but limited-rate data service. However, WLANs support high-rate data service, but lack of universal roaming ability and suffer from low QoS level for voice service, due to their original real-time constraints. Furthermore, call admission control has been implemented in cellular network to ensure low call dropping probability in system by assigning voice horizontal handoffs with a higher priority for resource than new voice and data call requests, while WLANs only support coarse packet-level access without considering handoffs priorities. So in integrated WLAN and 3G cellular networks, seamless vertical handoffs and call admission control must be considered as dependent and joint mechanisms to ensure both high-level call service quality and efficient resource utilization in interworking environment.

3. Call admission control and previous work

In communication system, the call admission control scheme is a provisioning strategy for QoS provisioning and network congestion reduction (Ahmed, 2005). Arriving calls are granted or denied based on predefined system criteria. Due to limited spectrum resource and growing popularity of usage in wireless cellular networks, CAC has been receiving a lot of attentions for QoS provisioning, and its main features are extended to cover signal quality, blocking probability of new call, handoff dropping probability, data rate, etc. The next-generation integrated WLAN and 3G cellular networks pose a great challenge to the CAC design due to heterogeneous network features, such as varied access techniques, resource allocation priorities, QoS provisioning levels, vertical handoffs, etc.

3.1 Call admission control in cellular networks

Extensive research work has been done on the CAC schemes in homogeneous cellular networks (Ahmed, 2005). They can be classified based on various design focuses and algorithms, and each algorithm has its own advantages and disadvantages. Generally, CAC in 3G cellular networks give higher priority for voice service than data services for resource allocation, and higher priority for handoff calls than new call requests. We classify previous work on CAC into five major categories: signal quality based CAC, guard channel reservation based schemes, queuing methods, QoS estimation methods, and bandwidth degradation approaches.

Signal quality based CAC: signal quality in the physical layer is used as a criterion of admission control (Ahmed, 2005; Liu & Zarki, 1994). Some research work use power level of received signals or signal-to-noise-ratio (SIR) threshold as call admission requirements (Liu & Zarki, 1994). An optimal CAC scheme is proposed to minimize the blocking probability while keeping a good signal quality to reduce the packet error (Ahmed, 2005). However, all the above schemes only check the signal characteristics in the physical layer without considering technical features in other layers and service priorities. Furthermore, there are different criteria for the measurement of signal quality in integrated networks. So it is difficult for implement a CAC in an interworking environment based on a uniform criterion.

Guard channel reservation based schemes: To prioritize handoff calls over new calls, a number of channels, guard channels, in each cell are reserved for exclusive use by handoff calls, while the remaining channels are shared by both new calls and handoff calls. To decrease the handoff call dropping probability, which is at the cost of increasing the new call blocking probability, the guard channel must be chosen carefully and dynamically adjusted so that the dropping probability of handoff call is minimized and the network can support as many new call requests as possible (Fang & Zhang, 2002; Ahmed, 2005). However, the intensities of new call requests and handoff requests are time-variant, and it is difficult to assign appropriate guard channel timely. So the guard channel will reduce the efficiency of system resource utilization, and may not be suitable for heterogeneous network environment.

Queuing methods: When there is no channel for incoming call requests, either handoff call requests are put into a queue while new call requests are blocked, or new call requests are put into a queue while handoff calls are dropped (Lau & Maric, 1998; Ahmed, 2005). Although queuing schemes can avoid high blocking probability or dropping probability due to increased call intensity for a short period, it is not realistic in a practical system in which a handoff call may not hold in a queue for a long time because of fast signal fading, and new calls will leave the queue system due to users' impatience.

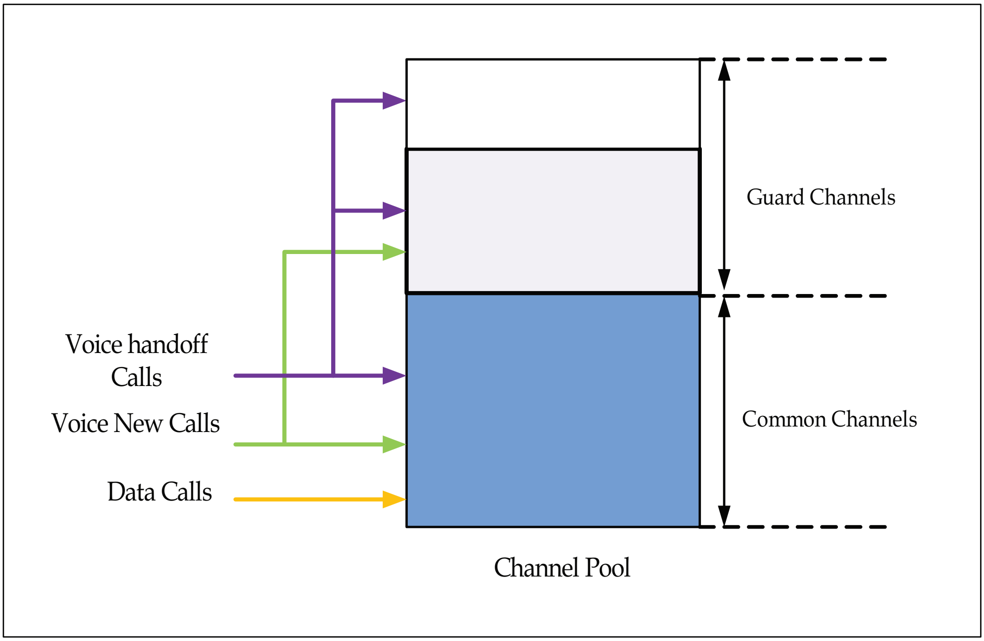


Fig. 3. Guard channel reservation based scheme for voice and data services

QoS estimation based approaches: CAC in cellular networks calculates the future resource requirements for new calls and handoff calls based on user mobility and call intensity estimation (Zhao et al., 2003; Koodli & Puuskari, 2001). A weighted overall handoff failure probability for all cells is calculated as an indicator for long-term statistics of successful call completion. The suggested schemes take the overall weighted handoff failure probability as the criterion for new call admission. Although those schemes can improve the efficiency of admission control and resource utilization, they cause nontrivial calculation complexity, and too many real-time control messages among cells may incur large signaling traffic and communication overhead. Furthermore, rough estimation techniques used in these schemes may cause erroneous decisions for call requests in a real world scenario, which will deteriorate the QoS level in the system.

Bandwidth degradation CAC: Some methods are proposed to degrade some connections adaptively when there are no more resources available for incoming new calls or handoff calls. For example, longest calls in the system are degraded to free resource for handoff calls (Jia & Mermelstein, 1996). Another proposal includes an algorithm in which each admitted connection degrades to a lower bandwidth level according to weights (Ahmed, 2005). Other proposals reduce the bandwidth of latest admitted connections. However, bandwidth degradation can only reduce the bandwidth of varied-bit-rate (VBR) and non-real-time (NRT) services for each individual, and is not suitable for constant-bit-rate (CBR) connections. Furthermore, though these schemes can reduce the blocking probability, the QoS level in the network cannot avoid deteriorating after degradation, and the overall utilization ratio may not be improved.

3.2 Call admission control in integrated WLAN and 3G cellular networks

There have been some works on call admission control in integrated WLAN and 3G cellulr networks. Most significant ones are WLAN-first approaches, mobility based algorithms and policy based CAC schemes.

WLAN-first approaches: If mobile terminals locate in a WLAN service area, both new voice and data calls first request admission to the WLAN. If rejected, the calls overflow to 3G cellular network. If mobile terminals with on-going voice and data calls move into the WLAN, the calls always try to handoff to WLAN (Song et al., 2006; Song et al., 2007a). This unconidional preference to WLAN aims to take advantage of cheaper and higher bandwidth in WLAN, compared to 3G cellular network. However, these approaches may cause an over-crowded traffic situation in WLAN, without load balance in both networks. Since user mobility is not covered in these approaches, frequent handoff requests will happen around the WLAN boundary, which may cause "Ping-Pong" effects for too many vertical handoffs with extra large sigaling traffics generated into networks.

Mobility based algorithms: Some research works consider users with different mobility speeds and apply different CAC and vertical handoff algorithms for them. Some authors probabilistically reject vertical handoff requests to WLAN for highly mobile cellular users (Lampropoulos et al., 2005; Klein & Han, 2004), to reduce unnecessary handoffs. In this scheme, the processing load and new call blocking probability can be reduced while maintaining reasonable throughput in the WLAN. A mobility-based predictive call admission control technique has been proposed for the 4G wireless heterogeneous networks (Rashad, 2006). In this scheme, local and global mobility profiles for the mobile terminals are generated and used for call admission decision. However, since randomness of user mobility, it may be difficult for such algorithms to getting speed estimation timely and concisely. Besides handoff management based on mobility information, more works are needed for considering service differentiations, QoS cost, and user preference, to provide global optimization for resource utilization in integrated networks.

Policy Based CAC Schemes: Some solutions follow policy framework defined by IETF, and combine call admission control and vertical handoff management together. They use a mobile-assisted scheme, in which system functionality is controlled by a network policy engine and a mobile policy engine (Zhuang et al., 2003). As shown in Fig. 4, a pairing of a policy decision point (PDP) and policy enforcement point (PEP) exist in both engines, along with policy repositories. PEP is responsible for the execution of a policy that is decided by PDP, and the policy repositories define the policies that must be followed for a proper handover decision (Zhuang et al., 2003; Guerrero & Barba 2008). In the call admission control procedure, PEPs in the mobile terminals consult a PDP residing at the network for available resources. The PDP will make a decision on call admission, based on network capacities, QoS level, call types, user preferences as well as estimations on current network load and performances. This approach gives flexibility to the terminal and the network to make the best possible handover decision, and implements load balance. However, there are several drawbacks of this policy method, such as high latencies to fetch context information during the candidate access point classification procedure, and no optimization policy is defined for resource allocation in integrated networks.

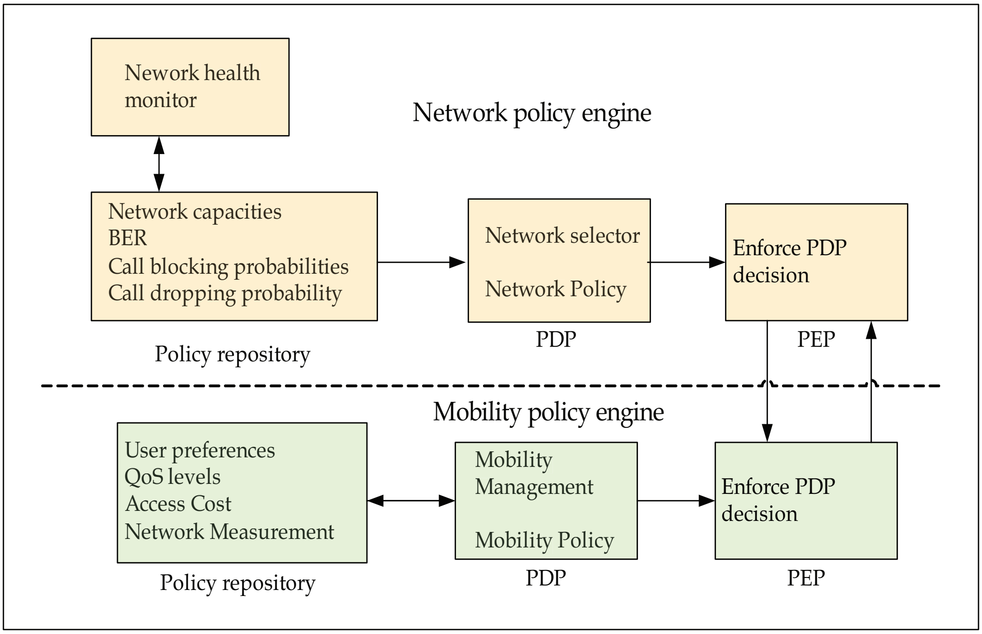


Fig. 4. Policy-based admission control framework

4. A novel joint call admission control scheme

In this section, a novel joint call admission control scheme is proposed for both voice and data services with QoS provisioning in integrated WLAN and 3G UMTS cellular networks. Due to different network service characteristics, UMTS network is defined to be a Voice-Priority network where voice services have higher priority for resource allocation than data services, while WLAN is defined as Data-Priority network where data services have higher priority than voice services. Instead of using WLAN-first schemes, a novel joint call admission policy is derived to support heterogeneous network architecture, service types, QoS levels, mobility characteristics, and user network preferences. Based on the policy, a channel searching and replacement algorithm (CSR) is designed to relieve traffic congestion in UMTS cellular network. CSR searches idle channels in WLAN and replacement channels among mobile terminals based on their location and passive vertical handoffs, and therefore implements load balance between UMTS networks and WLAN. A one-dimension Markov model is further developed to compare system performance metrics, such as new call blocking probability and dropping probability of voice handoff, between the proposed algorithm and normal disjoint guard channel reservation based scheme.

The CSR algorithm is further improved by considering congestion scenarios in both the WLAN and UMTS cellular networks. Specifically, a system cost function is derived and minimized by admitting passive vertical handoffs with a probability, and it is proven that there exists at least one optimal value for the target passive handoff probability. In this way, the total traffic is balanced in the interworking environment as well as the resource utilization is optimized. A linear programming solution is proposed for searching the optimal admission probability for passive vertical handoff, with performance comparsion to the traditional disjoint guard channel CAC. Numerical and simulation results demonstrate that the optimal CSR (oCSR), outperforms disjoint guard channel CAC and original CSR algorithms in both QoS provisioning and system resource utilization.

4.1 System model

Our system model is based on the loose coupling achitecture in which WLAN is connected to cellular networks through Internet. All traffics in WLAN are routed to Internet through gateway routers. Since the coverage area of a UMTS cell is normally much larger than WLAN area, the cellular cell is called a "macro-cell," while the WLAN region is regarded as a "micro-cell" inside (Liu et al., 2007). The overlaid service area between the "macro-cell" and the "micro-cell" provides mobile terminals with opportunities to connect to either UMTS network or the WLAN, as shown in Fig. 5.

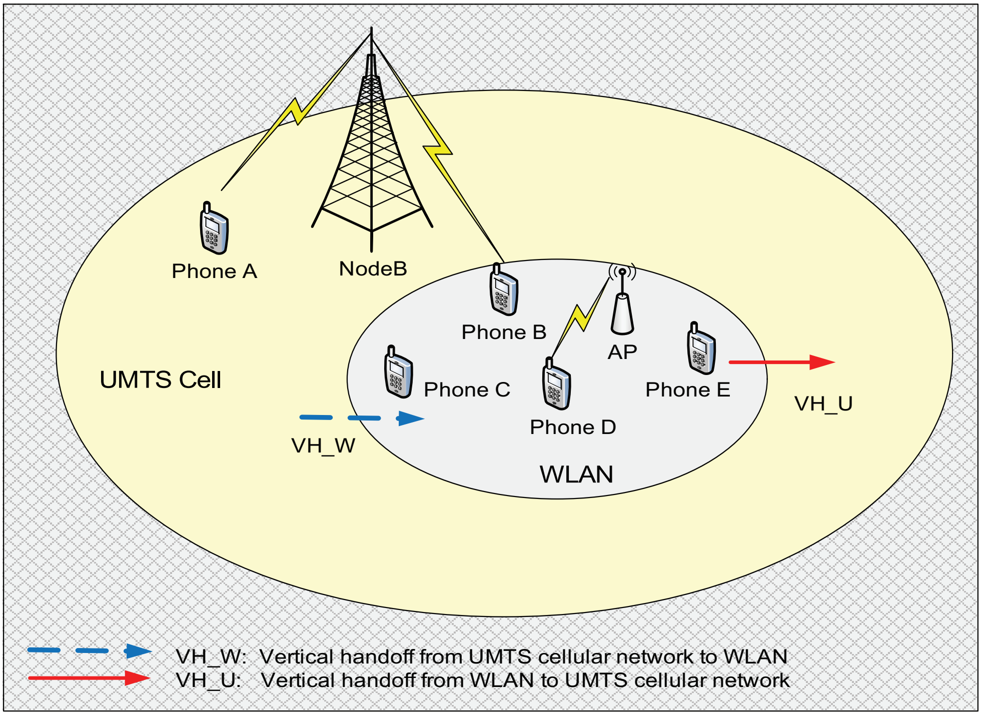

Phone A
NodeB
Phone B
AP
UMTS Cell
Phone C
Phone D
Phone E
VH_U
VH_W
WLAN
VH_W: Vertical handoff from UMTS cellular network to WLAN
VH_U: Vertical handoff from WLAN to UMTS cellular network

Fig. 5. System model for integrated UMTS cell and WLAN

There are one class of voice service and *N* classes of data service considered in the integrated networks. Since voice is sensitive to the transmission delay, the QoS requirement for voice service is represented as an average voice packet generation rate b_v and the maximum tolerable transmission delay D_v. On the other hand, QoS constraint for data service *i* is mainly defined as the minimum throughput *Ti*, since data services are assumed as best effort services and not sensitive to the transmission delay.

With careful network planning and mature admission control algorithms, the UMTS network is assumed to support both voice and data services with QoS provisioning. However, due to limited bandwidth and high cost per bit, voice calls are assigned with higher priority than data services in cellular networks, and data services are regarded as best-effort services. So we set the proportion *Rc* (0< *Rc* < 1) as the maximum ratio of resource usage assigned to all data services in the cellular network.

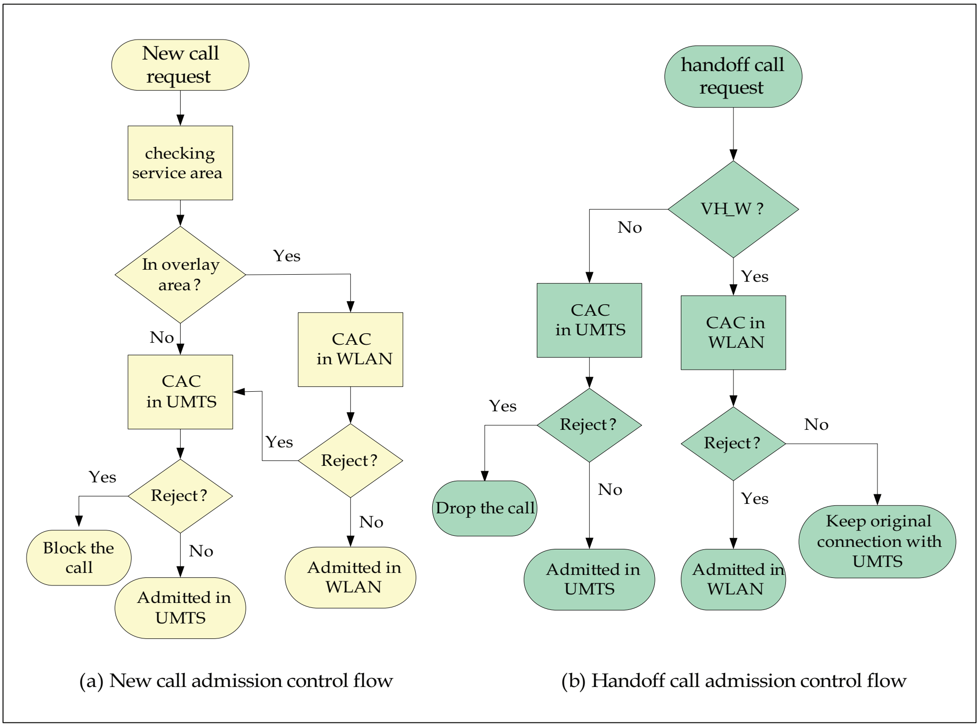
On the other hand, the traditional WLAN is designed to support only best-effort data services. To match the QoS difference between the cellular network and the traditional WLAN, we assume that the IEEE 802.11e WLAN standard is adopted and our polling algorithm named IIT (Liu & Zhou, 2006) is used to guarantee bandwidth and delay requirements for voice services in WLAN while enhanced distributed coordination function (EDCF) to support data services. Since normally mobile users connect to WLAN for high-speed data transmission with low cost, we assume that data services in WLAN have higher priority than voice services. Compared to cellular network, the proportion *Rw* (0 < *Rw* < 1), is set as the maximum ratio of resource usage assigned to all voice services in the WLAN.

Based on user mobility characteristics, voice call requests in a UMTS cell can be classified into the following two categories: vertical handoffs from the overlaid WLAN to the UMTS, which is denoted as *VH_U*, and new calls from this UMTS cell. Since a WLAN network is assumed to be overlaid by a UMTS cell, there are also two categories in admission requests: new call requests for WLAN access, and vertical handoffs from the surrounding UMTS cell, denoted as *VH_W*, as shown in Fig. 5. To reduce complexity, the horizontal handoffs from nearby cellular cells to the target cell are assumed to be equal to horizontal handoffs out of the cell to neighbor cells so that we do not consider horizontal handoffs in this work. We further assume that QoS level in the integrated networks are mainly bottlenecked by the total bandwidth in the integrated networks. Therefore, the QoS performance considered in this work includes blocking probability of new call requests and dropping probability of handoff call requests.

4.2 Call admission control flow and policy

The admission control flow for the integrated networks is classified into two sub-schemes according to the types of call requests, as shown in Fig. 6. If the call request is a new call and mobile terminal is out of the WLAN service area, the mobile terminal will send the new call request to the UMTS base station directly. On the other hand, if the call is a new call and mobile terminal is in the WLAN area, the mobile terminal will first send a new call request to the WLAN as the first choice, since the WLAN is much less expensive per call. The new call request will be handled by the CAC in the WLAN. If the WLAN cannot accommodate the new call request, the request will be forwarded to the UMTS base station.

The situation becomes complex when we consider vertical handoff call requests. Since voice service is real-time service, a user is much more sensitive to voice dropping than dropping of data service during vertical handoff. So voice vertical handoff from WLAN to UMTS cellular network should be assigned a higher priority than new call requests and data vertical handoff from WLAN to UMTS, to avoid possible voice dropping in the cellular service area. Here, we adopt guard channel scheme (Fang & Zhang, 2002) to reserve some bandwidth for voice *VH_U* handoff. However, for any data *VH_U* handoff, no matter how high data rate it gets in WLAN, it become best-effort service when the user move from WLAN to cellular network. So data *VH_U* is assigned same priority as any new call requests is in cellular network.

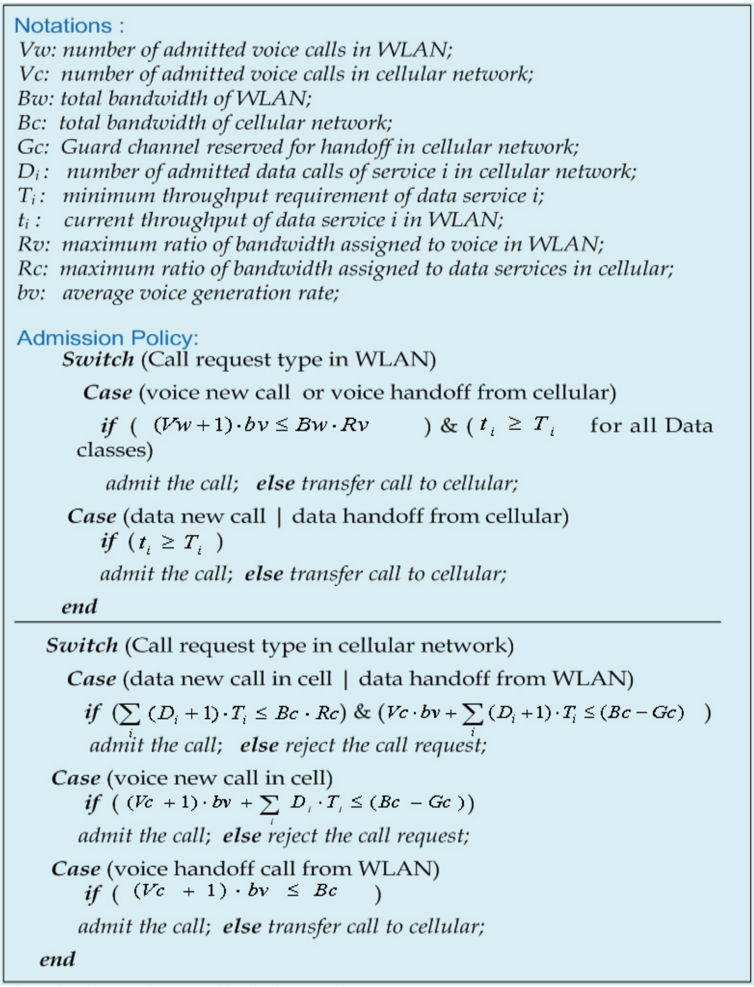


(a) New call admission control flow (b) Handoff call admission control flow

Fig. 6. Call admission control flow

Furthermore, for all *VH_W* handoff requests from cellular to WLAN, users can still keep original connections with the cellular network even when the handoff request is denied by the WLAN. Since there is no real connection dropping for *VH_W* handoffs, both *VH_W* and new calls in the WLAN are assigned the same priority in our system.

Based on the above analysis of user mobility and service characteristics, the pseudocode of proposed admission policy for interworking system are developed in Fig. 7. We assume total bandwidth in cellular network and WLAN as *Bc* and *Bw*, respectively. The numbers of admitted voice call in WLAN and cellular network are denoted as *Vw* and *Vc*, respectively. The guard channel in cellular network is set as *Gc*. The number of admitted data service of class-*i* is denoted as *Di*. When WLAN receive a call request, the average throughput, t_i, for current WLAN can be calculated based on enhanced DCF mode method (Liu & Zhou, 2006).



Notations :

Vw: number of admitted voice calls in WLAN;
Vc: number of admitted voice calls in cellular network;
Bw: total bandwidth of WLAN;
Bc: total bandwidth of cellular network;
Gc: Guard channel reserved for handoff in cellular network;
D_i: number of admitted data calls of service i in cellular network;
T_i: minimum throughput requirement of data service i;
t_i: current throughput of data service i in WLAN;
Rv: maximum ratio of bandwidth assigned to voice in WLAN;
Rc: maximum ratio of bandwidth assigned to data services in cellular;
bv: average voice generation rate;

Admission Policy:

```
Switch (Call request type in WLAN)
    Case (voice new call  or voice handoff from cellular)
        if ( $(Vw+1)\cdot bv \le Bw\cdot Rv$ ) & ( $t_i \ge T_i$ for all Data classes)
            admit the call;  else transfer call to cellular;
    Case (data new call | data handoff from cellular)
        if ( $t_i \ge T_i$ )
            admit the call;  else transfer call to cellular;
end
```

```
Switch (Call request type in cellular network)
    Case (data new call in cell | data handoff from WLAN)
        if ( $\sum_i (D_i+1)\cdot T_i \le Bc\cdot Rc$ ) & ( $Vc\cdot bv + \sum_i (D_i+1)\cdot T_i \le (Bc-Gc)$ )
            admit the call;  else reject the call request;
    Case (voice new call in cell)
        if ( $(Vc+1)\cdot bv + \sum_i D_i\cdot T_i \le (Bc-Gc)$ )
            admit the call;  else reject the call request;
    Case (voice handoff call from WLAN)
        if ( $(Vc+1)\cdot bv \le Bc$ )
            admit the call;  else transfer call to cellular;
end
```

Fig. 7. Call admission control policy

4.3 Channel searching and replacement (CSR) algorithm

Although the above proposed CAC can handle call requests in both WLAN and cellular networks, all admission decisions are made based on the situation of each individual network. To improve the whole system performance, we propose a channel searching and replacement (*CSR*) algorithm based on passive vertcial handoff to implement joint resource management.

Due to different capacities and user densities, the traffic intensities and QoS levels are often unbalanced in the WLAN and overlaid cellular network. When WLAN becomes congested, the traffic will be routed to the cellular network automatically. On the other hand, when the 3G cellular network has no resource available for an incoming call requests, our CSR algorithm is used to find available resources in the WLAN by switching some 3G connections staying in WLAN area to the WLAN, as shown in Figure 8. Specifically, if there exists an ongoing cellular connection and the mobile terminal residing in the WLAN area, and there is still bandwidth available in the WLAN at the same time, the cellular connection will be switched to the WLAN by vertical handoff, and then the incoming call request will take the released bandwidth in cellular network to avoid being blocked or dropped. This kind of vertical handoff is called "passive" because it is initiated by the system resource management instead of by users or signal fading.

To achieve the fairness among different service connections, *CSR* checks the difference of QoS provisioning in both networks before switching a cellular connection to WLAN. If there is no QoS degradation during switching and WLAN can guarantee QoS provisioning for all existing ongoing calls, then the bandwidth or channel is released.

Considering the CSR algorithm may increase the blocking probability in the WLAN (i.e., deteriorate the QoS in the WLAN by forwarding more traffics from the cellular network to WLAN). We further assume that there is a call admission probability for passive vertical handoff, which is determined by the system status of cellular network and WLAN, and QoS levels. The pseudocode of the *CSR* is shown in Fig. 8.

```
switch (call request in cellular network)
case (data-call-arrival):
    if (CAC for data::admitted) & (QoS provisioning )
        admit the call;
    else if (Channel_Searching() == 1) & (No degradation)
        switch the cellular connection to WLAN;
        admit the call request & assign a channel with a probability P;
    else { reject the call request;}
        break;
case (voice-call-arrival):
    if (CAC for voice::admitted) & (QoS provisioning )
        admit the call ;
    else if (Channel_Searching() == 1) & (No degradation)
        switch the cellular connection to WLAN;
        admit the call request & assign a channel with a probability P;
    else { reject the call request; }
        break;
default: break;
end
-------------------------------------------------------------------
#Channel_searching() :
Search for cellular connections but mobile terminal staying in WLAN;
if (at least one cellular connection in WLAN) & (QoS provisioning in
WLAN ) { return 1; }
else { return 0; }
```

Fig. 8. Channel searching and replacement (CSR) algorithm

4.4 Analysis and comparsion

In this section, the proposed CSR algorithm is compared with traditional disjoint guard channel (DGC) scheme with system performance metrics, including new call blocking prabability and handoff dropping probability. To reduce the complexity, we focus on voice services in the integrated WLAN and 3G UMTS cellular networks, with fixed total channels in UMTS cell and bandwidth in WLAN.

4.4.1 DGC algorithm

First the traditional DGC algorithm is considered. Assume that the arrival process for both new calls and vertical handoff follows Poisson distributions, and the channel holding time for both vertical handoffs and new calls are exponentially distributed. Let λ_n and $1/\mu_n$ denote the arrival rate and the average channel holding time for new voice call in the UMTS cell, respectively. Let λ_v and $1/\mu_v$ denote the the arrival rate and average channel holding time for voice vertical handoff from WLAN to UMTS cell, respectively. The arrivals of new calls and vertical handoffs are independent of each other. To simplify, assume the avarage channel holding time for both new voice call and handoff call are same: $\mu_n = \mu_v$.

Assume total C available channels in UMTS cellular network for voice service. An approximate one-dimension Markov model (Fang & Zhang, 2002; Liu et al., 2007) is derived to present state transitions in UMTS network, as shown in Fig. 9(a). The state space in cellular network can be denoted as $\{(m,n) \mid 0 \le m+n \le C\}$, where m and n are the numbers of admitted new calls and admitted vertical handoffs in the cell, respectively. The traffic intensity of vertical handoffs ω_v and traffic intensity of new calls ω_n are specified as $\omega_v = \lambda_v / \mu_v$ and $\omega_n = \lambda_n / \mu_n$, respectively.

Based on the stationary state distribution, the vertical handoff dropping probability P_v and new call blocking probability P_n , for disjoint guard channel scheme can be expressed as follows,

$$P_v = \pi_c(C) = \frac{\dfrac{(\omega_n+\omega_v)^G \cdot (\omega_v)^{C-G}}{C!}}{\sum\limits_{i=0}^{G}\dfrac{(\omega_n+\omega_v)^i}{i!} + \sum\limits_{i=G+1}^{C}\dfrac{(\omega_n+\omega_v)^G (\omega_v)^{i-G}}{i!}} \tag{1}$$

$$P_n = \sum_{i=G}^{C}\pi_c(i) = \frac{\sum\limits_{i=G}^{C}\dfrac{(\omega_n+\omega_v)^G \cdot (\omega_v)^{i-G}}{i!}}{\sum\limits_{i=0}^{G}\dfrac{(\omega_n+\omega_v)^i}{i!} + \sum\limits_{i=G+1}^{C}\dfrac{(\omega_n+\omega_v)^G (\omega_v)^{i-G}}{i!}} \tag{2}$$

where $\pi_c(i)$ represents the stationary state of occupied channel i. The detailed derivations for above equations are shown in our previous work (Liu & Zhou, 2007).

4.4.2 CSR algorithm

In the proposed CSR scheme, the total number of occupied channels in the cell and the idle channels in the WLAN are the keys to deciding whether a new voice calls or a vertical handoffs need intersystem channel switching through a passive handoff to the WLAN. When the total channel number i in the cell is larger than *Gc*, an incoming new call request can get admission if there is an ongoing cellular connection residing the WLAN and there is still bandwidth available in the WLAN. When the total occupied UMTS channel number equals to *C*, an incoming vertical handoff from WLAN can also be admitted in cellular network if there is a successful channel replacement in the WLAN. To avoid over-utlization on WLAN, it is assumed that a call request can get admission with probability δ that is determined by the total number of occupied channels in the cell, the probability for mobile terminals using ongoing cellular connection while located in the WLAN, and the state of current occupied channels in the WLAN. Based on the above descriptions, we can get a Markov chain model for the cellular network, shown in Fig 9(b).

Using CSR, call request blocking or dropping in a cellular network will happen in following two scenarios:

Scenario 1: There is no idle channel available in cellular network, and no cellular connections residing in the WLAN;

Scenario 2: There is no idle channel available in cellular network, and no channel within the WLAN, although there is a cellular connection residing in the WLAN.

So Let P_f be the probability of an ongoing cellular call remaining in a WLAN, which is assumed to be determined by a user's preference for vertical handoff and mobility velocity. Let $\psi_c(i)$ be the probability that there is no cellular connection within the WLAN when the number of total occupied channels in the cellular network is *i*.

$$\psi_c(i) = \binom{i}{0} \cdot (p_f)^0 \cdot (1-p_f)^i \tag{3}$$

If the probability for finding a cellular connection staying in the WLAN is set as 1, which means always finding available cellular connection successfully, the traffic intensity in the WLAN depends on not only original traffic inside, but also on passive handoffs from the cell. So the traffic intensity $\rho(i)$ in the WLAN is a function of state *i* in UMTS cell and can be expressed as,

$$\rho(i) = I_1(i) \cdot (\rho_n + \rho_v) + I_2(i) \cdot (\rho_n + \rho_v + \omega_n) + I_3(i) \cdot (\rho_n + \rho_v + \omega_n + \omega_v) \tag{4}$$

where ρ_n is original traffic intensity of new call requests in WLAN, ρ_v is original call intensity of vertical handoff requests from UMTS to WLAN. $I_i()$ are state indicator functions: $I_1(i)$ equals to 1 when state *i* smaller than guard channel Gc, otherwise equals to zero. $I_2(i)$ equals to 1 when state *i* larger than Gc-1 and smaller than total channels *C* in UMTS cell, otherwise equals to zero. $I_3(i)$ equals to 1 when state *i* equals to total channels C in UMTS cell, otherwise equals to zero.

Since in WLAN vertical handoffs and new calls are assigned with same priorities for resource, the blocking probability of new call is same to dropping probability of vertical handoffs. Considering voice service, the blocking probability p_b^w in WLAN is determined by incoming traffic intensity $\rho(i)$, which is affected by traffic intensities in both UMTS cell and WLAN, the probability of an ongoing cellular call remaining in a WLAN, as well as admission probability of passive handoffs.

According to above definitions of the two scenarios, the blocking probability for new call requests and dropping probability for vertical handoffs from WLAN to cellular network can be approximated as,

$$P_n = \sum_{i=G}^{C} \left\{ \psi_c(i) + \left[1 - \psi_c(i)\right] \cdot p_b^w(i) \right\} \cdot \pi_c(i) \tag{5}$$

$$P_v = \left\{\psi_c(C) + \left[1-\psi_c(C)\right]\cdot p_b^w(C)\right\}\cdot \pi_c(C) \tag{6}$$

where $\pi_c(i)$ represents the stationary state of occupied channel *i* in UMTS cell.
Since probability that there is no cellular connection within the WLAN is alway smaller than 1, and same for blocking probability p_b^w in WLAN, it is proved (Liu & Zhou, 2007) that value of blocking probability for new call requests and dropping probability of vertical handoffs in UMTS cell through CSR algorithm are both smaller than the probability values using disjoint guard channels shown in equations (1) and (2).

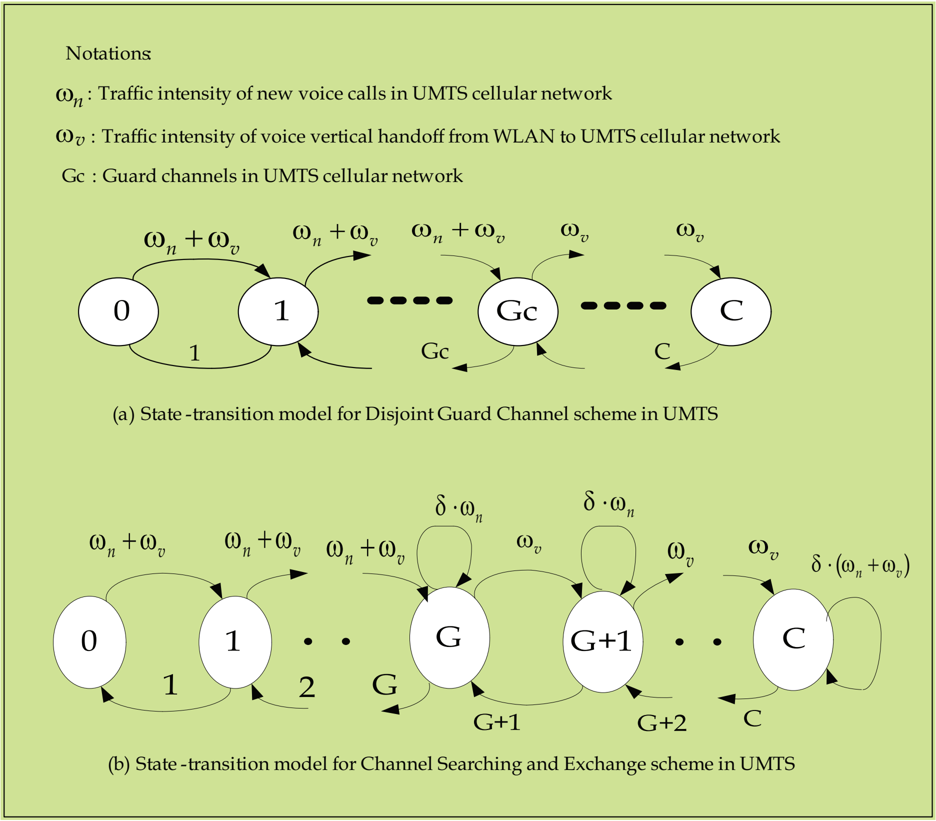

Notations:
ω_n : Traffic intensity of new voice calls in UMTS cellular network
ω_v : Traffic intensity of voice vertical handoff from WLAN to UMTS cellular network
Gc : Guard channels in UMTS cellular network
$\omega_n+\omega_v$
$\omega_n+\omega_v$
$\omega_n+\omega_v$
ω_v
ω_v
0
1
Gc
C
1
Gc
C
(a) State-transition model for Disjoint Guard Channel scheme in UMTS
$\delta\cdot\omega_n$
$\delta\cdot\omega_n$
$\omega_n+\omega_v$
$\omega_n+\omega_v$
$\omega_n+\omega_v$
ω_v
ω_v
ω_v
$\delta\cdot(\omega_n+\omega_v)$
0
1
G
G+1
C
1
2
G
G+1
G+2
C
(b) State-transition model for Channel Searching and Exchange scheme in UMTS

Fig. 9. State-transition diagram for DGC and CSR algorithms

4.5 Optimization on joint call admission control

Although the blocking probability of new calls and dropping probability of handoff calls in UMTS cellular network get reduced by using CSR algorithm, the cost is load balance traffics to WLAN and therefore may deteriorate QoS in WLAN, such as increasing blocking probability in WLAN. So the joint call admission control needs to be optimized to achieve the minimum blocking probability per Erlang in the integrated networks.
A weitghted system cost function is derived based on blocking probability, dropping probability, call intensities, and probability of passive vertical handoffs. Our goal is to

minimize average weighted system cost with constraint on probability of passive vertical handoffs, as shown in follows:

$$\textbf{Minimize } P_{ave} = \frac{W_1 \cdot P_n \cdot \omega_n + W_2 \cdot P_v \cdot \omega_v + W_3 \cdot P_b^w \cdot (\rho_n + \rho_v)}{\omega_n + \omega_v + \rho_n + \rho_v}$$

s.t. $0 \le \delta \le 1$

where W_1, W_2, and W_3 are cost weights for the blocking probability in the cellular network, the dropping probability in cellular network, and the blocking probability in the WLAN, respectively.

It is easy to prove that blocking probability in WLAN is a monotonically increasing continuous function of δ, while blocking probability and dropping probability in UMTS cell are continuous decreasing functions over δ in the interval between zero and one. So the weighted cost function is also a continuous function over the same interval. According to the Extreme Value Theorem, target cost function has a minimum and a maximum value over the interval $0 \le \delta \le 1$. So it is feasible to find out a optimal admission probability for passive handoff which minimizes the integrated system cost with linear programming. Here we should notice that there may be more than one optimal value for the admission probability.

5. Numerical and simulation results

In this section, the performances of CSR are testified through numerical results and simulations. Referred from (Fang & Zhang, 2002; Liu, 2006; Liu et al., 2007), the system parameter values are shown in Table 1, and results are shown as below. We focus on voice service and assume that the traffic intensity of data service in both WLAN and cellular network are kept constant. The step searching method of linear programming (Liu, 2006) is used to find the optimal admission probability for passive vertical handoff.

Bc	Bw	Gc	bv	Rc	Rw	p_f	W_1	W_2	W_3	Ti
20	30ms	18	30kb	0.2	0.2	0.3	1.0	2.0	1.0	30kb

Table 1. System parameters

Fig. 10 shows the changes in the optimal admission probability for passive vertical handoff as handoff intensity in the cell varies. We set new call intensity in UMTS cell ω_n = 10, new call intensity in WLAN ρ_n = 10, vertical handoff intensity ρ_v = 5. Since the weight of handoff dropping is larger than both the weights of blocking calls in cellular network and in WLAN, the optimal admission probability increases quickly for W3 = 1.3 and W3 = 2.0, and is 1 when the handoff intensity is larger than 45. In other words, the integrated system attempts to allocate each idle resource in the WLAN to handoff in cellular network to avoid larger system cost caused by dropping probability.

In contrast, when new call intensity ρ_n in the WLAN increases (ω_v is set as 5), the admission probability for W3 = 2.0 and W3 = 1.3 is reduced to zero, but remains 1 for W3 = 1, as shown in Figure 11. Again, it is shown that CSR can adjust the traffic intensity among the two networks to avoid overloaded situation in the WLAN. For W3 = 1.0, since the cost for blocking a passive handoff is no more than the costs of blocking a new call or dropping a connection in cellular network, the passive handoff always get an admission into the WLAN.

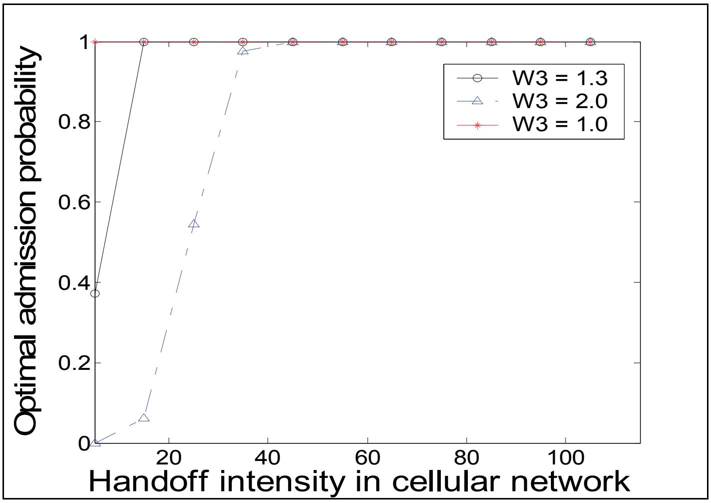


Fig. 10. Optimal admission probability for passive handoff vs handoff intensity in cellular

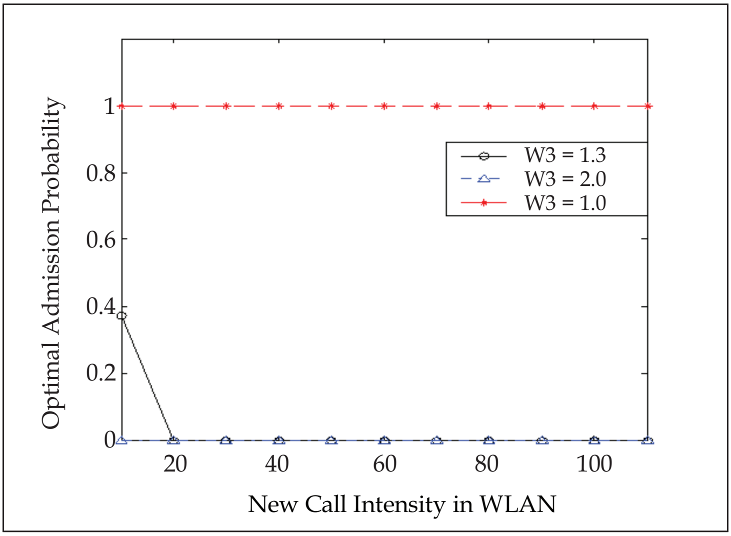


Fig. 11. Optimal admission probability for passive handoff vs new call intensity in WLAN

To validate the analytical results, simulations were performed based on the OPNET tool, an efficient discrete event-driven simulator. Fig. 12 shows the average system cost for DGC, CSR, and optimal CSR (oCSR), when new call intensity in UMTS, ω_n, is set as 30. In this case, the optimal admission probibility for passive handoff δ can be obtained as 0.078. DGC has the highest system cost due to its disjoint resource allocation, while oCSR can achieve the optimal resource allocation with minimum average system cost. Since the cost of oCSR is less than that of CSR, original CSR in UMTS cellular network is a sub-optimal solution for the overall resource allocation for integrated networks.

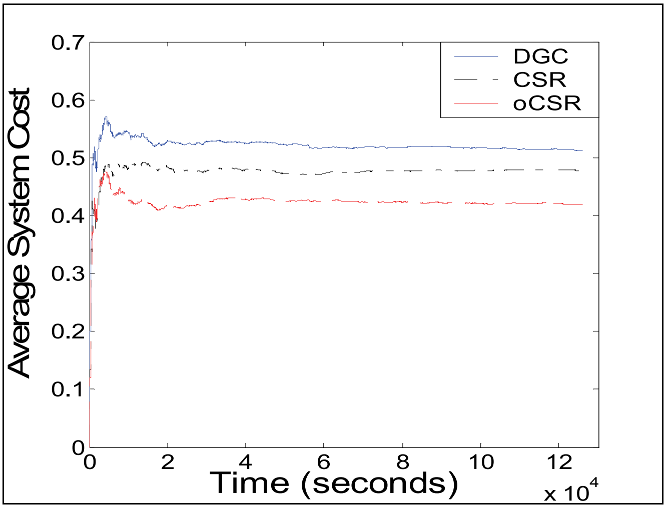


Fig. 12. System cost of DGC, CSR, and optimal CSR

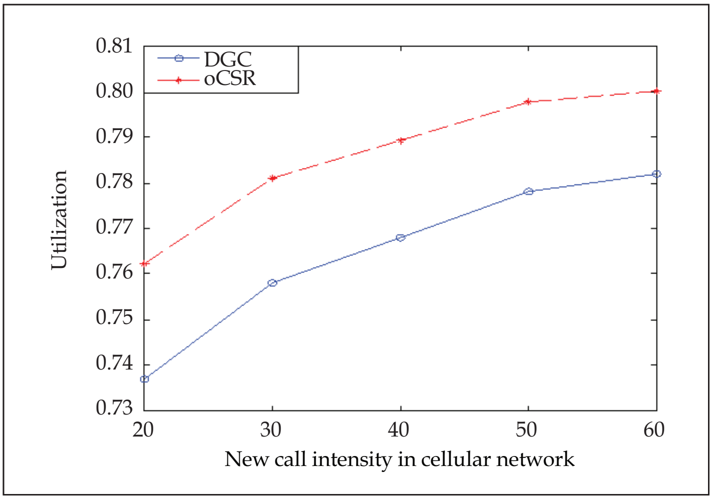


Fig. 13. Utilization with new call intensity in UMTS

Similarly, Fig. 13 shows the simulation result of utilization of system resource as new call requests ω_n in cellular network increases. We can see that optimal CSR has larger resource utilization than DGC does because optimal CSR uses idle resource in each network when traffic intensity in a network increases.

Fig. 14 shows the blocking probability when new call intensity in cellular network increases. When ω_n equals 20, 30, 40, 50, and 60, the optimal admission probability for passive handoffs are 0.496, 0.302, 0.216, 0.167, and 0.136, respectively. It is shown that the blocking probability of new call of oCSR scheme is always less than in the DGC scheme, due to optimal passive handoffs in oCSR scheme.

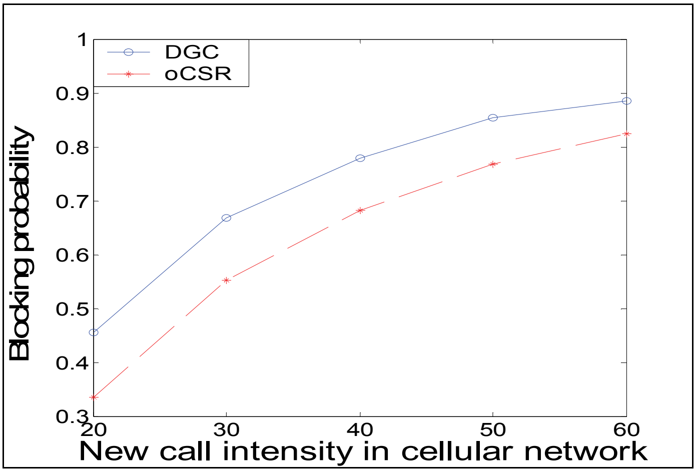


Fig. 14. Blocking probability with optimal CSR and DGC

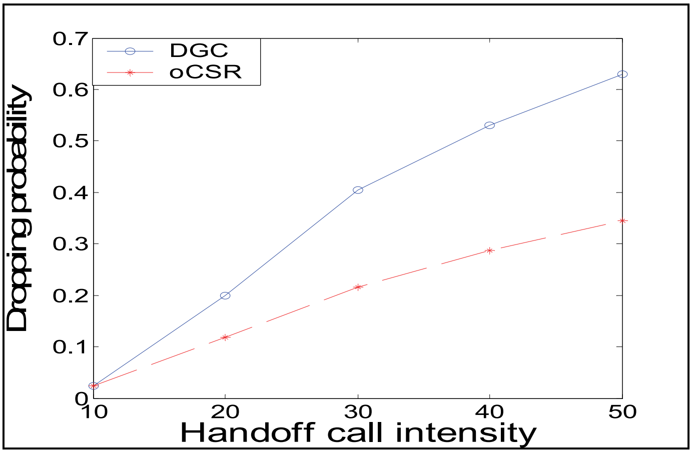


Fig. 15. Dropping probability with optimal CSR and DGC

Similarly, Fig. 15 shows the handoff dropping probability in the cell as the handoff intensity increases. Due to limited resources in the cellular network, both dropping probabilities increase. However, the dropping probability of the DGC is always greater than the dropping probability of the oCSR, since some handoffs are transferred to the WLAN, except in the case vertical handoff equals to 10. Since the optimal admission probability is equal to zero when $\omega_v = 10$, there is no passive handoff from the cellular network to the WLAN and both dropping probabilities are the same.

6. Conclusion

In this chapter, we introduce the next-generation call admission control schemes in integrated WLAN / 3G cellular networks. Technical background and previous works on call admission control in homogeneous and heterogeneous networks are investigated. Then a novel joint call admission control scheme is proposed to support both voice and data services with QoS provisioning in next-generation integrated WLAN / 3G UMTS networks. A joint admission policy is first derived with considering heterogeneous network architecture, service types, QoS levels, and user mobility characteristics. To relieve traffic congestion in networks, a channel searching and replacement algorithm, CSR, is further developed and optimized to balance total system traffics between WLAN and 3G cellular network, as well as to reduce average system QoS cost. A one-dimensional Markov model for voice traffic is further developed to analyze interworking system performance metrics. Both theoretical analysis and simulation results show that our scheme outperforms both traditional disjoint guard channel scheme and non-optimized joint call admission control scheme.

Our feature work will focus on more real-time services, such as video services, and investigate interactions between resource management and user mobility in integrated WLAN / 3G cellular networks.

7. References

Ahmavaara, K.; Haverinen, H. & Pichna, R. (2003), Interworking Architecture between 3GPP and WLAN Systems, *IEEE Communications Magazine*, Vol. 41, No.11, (Nov 2003), pp. 74 – 81, ISSN 0163-6804

Ahmed, M. (2005), Call admission control in wireless networks: a comprehensive survey, *IEEE Communications Surveys & Tutorials,* Vol. 7, No. 1, May 2005, pp. 50-69, ISSN 1553-877X.

Fang, Y. & Zhang, Y. (2002), Call admission control schemes and performance analysis in wireless mobile networks, *IEEE Transactions on vehicular Technology*, vol. 51, No.2, (March 2002), pp. 371-382, ISSN 0018-9545

Guerrero, J. & Barba, A. (2008), Policy-based Network Management Reference Architecture for an Integrated Environment WLAN-3G, *IEEE Latin America Transactions,* Vol. 6, No. 2, (June 2008), pp. 229-234, ISSN 1548-0992

Jia, D. & Mermelstein, P. (1996), Adaptive Traffic Admission for Integrated Services in CDMA Wirelessaccess Networks, *IEEE Journal on Selected Areas in Communications*, Vol. 14, No. 9, (Dec. 1996), pp.737–47, ISSN 0733-8716.

Klein, T. & Han, S. (2004), Assignment strategies for mobile data users in hierarchical overlay networks: Performance of optimal and adaptive strategies, *IEEE Journal on Selected Areas in Communications*, vol. 22, No. 5, (June 2004), pp. 849–861, ISSN 0733-8716

Koodli, R. & Puuskari, M. (2001), Supporting packet-data QoS in next generation cellular networks, *IEEE Communication Magazine*, Vol.39, No.2, (Feb. 2001), pp.180-188, ISSN 0163-6804

Lampropoulos, G.; Passas, N. & Merakos, L. (2005), Handover management architectures in integrated WLAN / Cellular Networks, *IEEE Communications Surveys & Tutorials,* Vol. 7, No. 4, (May 2005), pp. 30-44, ISSN 1553-877X.

Lau, K. & Maric, S. (1998), Mobility of Queued Call Requests of a new call-queueing technique for cellular systems, *IEEE Transactions on vehicular Technology*, vol. 47, no.2, (May 1998), pp.480-488, ISSN 0018-9545

Liu, C. & Zhou, C. (2004), Challenges and Solutions for Handoff Issues in 4G Wireless Systems An Overview, *Proceedings of International Latin American and Caribbean for*

Engineering and Technology 2004, Paper No. 047, Miami, Florida, USA, June 2-4, 2004.

Liu, C. & Zhou, C. (2005), HCRAS: A novel hybrid internetworking architecture between WLAN and UMTS cellular networks, *Proceedings of IEEE Consumer Communications & Networking Conference 2005*, pp. 374-379, ISBN 0-7803-8784-8, Las Vegas, Nevada, January 1- 10, 2005.

Liu, C. & Zhou, C. (2005), An improved architecture for UMTS-WLAN Tight Coupling, *Proceedings of IEEE Wireless Communications & Networking Conference 2005*, pp. 1690-1695, ISBN 0-7803-8966-2, New Orleans, LA, March 13-17, 2005.

Liu, C. & Zhou, C. (2006), Providing Quality of Service in IEEE 802.11 WLAN, *Proceedings of IEEE Consumer Communications & Networking Conference 2005*, pp. 374-379, ISBN 0-7803-8784-8, Las Vegas, Nevada, USA, January 1-10, 2006.

Liu, C. (2006). System Design and Resource Management in the Next-Generation Integrated WLAN / 3G Cellular Networks, *Doctoral dissertation*, Florida International University, Miami, Florida, USA, August 2006.

Liu, C.; Zhou, C.; Pissinou, N. & Makki, K. (2007), Resource Management in the Next-Generation Integrated WLAN / 3G Cellular Networks, *Proceedings of IEEE Wireless Communications & Networking Conference 2007*, pp. 3343-3348, ISBN 1-4244-0658-7, Hong Kong, China, March 11-15, 2007.

Liu, Z. & Zarki, M. (1994), SIR-based Call Admission Control for DS-CDMA Cellular Systems, *IEEE Journal on Selected Areas in Communications*, Vol.12, no. 4, (May 1994), pp. 638–44, ISSN 0733-8716

Rashad, S. (2006), Mobility-based predictive call admission control and resource reservation for next-generation mobile communications networks, Doctoral dissertation, University of Louisville, In ACM digital library, Available from
http://portal.acm.org/citation.cfm?id=12929

Shafiee, K. ; Attar, A. & Leung, V. (2011), Optimal Distributed Vertical Handoff Strategies in Vehicular Heterogeneous Networks, *IEEE Journal on Selected Areas in Communications,* Vol. 29 , No.3, (March 2011), pp. 534 – 544, ISSN 0733-8716

Song, W.; Jiang, H. & Zhuang, W. (2007), Performance analysis of the WLAN first scheme in cellular/WLAN interworking, IEEE *Transactions on Wireless Communications*, vol. 6, No. 5, (May 2007), pp. 1932-1943, ISSN 1536-1276

Song, W.; Jiang H.; Zhuang W. & Saleh, A. (2006), Call Admission Control for Integrated Voice/Data Services in Cellular/WLAN Interworking, *Proceedings of IEEE International Conference on Communications 2006*, PP. 5480 - 5485, ISBN 1-4244-0355-3, Istanbul, TURKEY, June 11-15, 2006

Song W.; Zhuang W. & Cheng Y. (2007). Load balancing for cellular/WLAN integrated networks, *IEEE Network,* Vol. 21, No. 1, (Feb 2007), pp. 27-33, ISSN 0890-8044.

Zhao, D.; Shen, X. & Mark, J. (2003), Radio Resource Management for Cellular CDMA Systems Supporting Heterogeneous Services, *IEEE transactions on Mobile Computing*, Vol.2, No.2, (June 2003), pp. 147 – 160, ISSN 1536-1233

Zhuang W.; Gan Y.; Loh K. & Chua K. (2003). Policy-based QoS-management architecture in an integrated UMTS and WLAN environment, *IEEE Communications Magazine,* Vol. 41, No. 11, (Nov. 2003), pp. 118-123, ISSN 0163-6804

10

Near-Optimal Nonlinear Forwarding Strategy for Two-Hop MIMO Relaying

Majid Nasiri Khormuji and Mikael Skoglund
Royal Institute of Technology (KTH)
Sweden

1. Introduction

Relaying (1–3) has been considered as a paradigm for improving the quality of service (i.e., bit-error-rate, data rate and coverage) in wireless networks. In this work, we study a two-hop relay channel in which each node can have multiple antennas. It is well-known that utilizing multiple-input multiple-output (MIMO) links can significantly improve the transmission rate (see e.g. (4; 5) and references therein). Thus, one can expect a combination of a MIMO gain and a relaying gain in a MIMO relay link. We focus on *one-shot* transmission, where the channel is used once for the transmission of one symbol representing a message. This is often referred to as *uncoded transmission*. The main motivation for such a scenario is in considering applications requiring either low-delays or limited processing complexity.

The capacity of the MIMO relay channel is studied in (6). The work in (9) establishes the optimal *linear* relaying scheme when perfect CSI is available at the nodes. The work in (7; 8) investigates *linear* relay processing for the MIMO relay channel. In this paper, in contrast to (6–9), we study an *uncoded* system, and we propose a *nonlinear* relaying scheme which is superior to linear relaying and performs close to the theoretical bound. Our proposed scheme is based on constellation permutation (10; 11) at the relay over different streams obtained by channel orthogonalization.

We investigate a two-hop MIMO fading Gaussian relay channel consisting of a source, a relay and a destination. We assume that all three nodes have access to perfect channel state information. We propose a nonlinear relaying scheme that can operate close to the optimal performance. The proposed scheme is constructed using channel orthogonalization by employing the singular value decomposition, and permutation mapping. We also demonstrate that linear relaying can amount to a significant loss in the performance.

1.1 Organization

The remainder of the chapter is organized as follows. Section 2 first introduces the two-hop relay channel model and then explains the transmission protocol and the assumptions on the channel state information (CSI) at the nodes and finally formulates an optimization problem. Section 3 simplifies and reformulates the optimization problem introduced in the preceding section, by channel orthogonalization using SVD. Section 4 introduces a novel relaying strategy in which the relay first detects the transmitted message and employs permutation coding over different streams obtained by channel orthogonalization. This section also

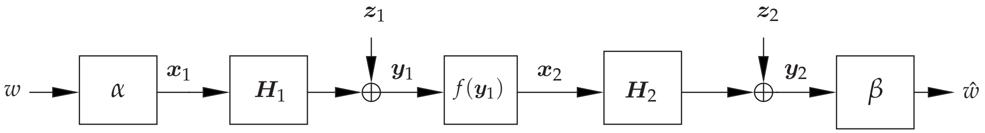


Fig. 1. Gaussian two-hop MIMO relaying.

provides some performance bounds. Section 5 finally provides some simulation results and concludes the chapter.

2. System model and problem formulation

In this section, we first introduce the two-hop Gaussian vector relay channel in detail and then formulate the general problem of finding an optimal relaying strategy for the underlying channel.

We consider Gaussian two-hop communication between a source and a destination, as illustrated in Fig. 1. The communication is assisted by a relay node located between the source and the destination. We assume that the relay node has no own information to transmit and its sole purpose is to *forward* the information received from the source to the destination. We additionally assume that all nodes may have different number of antennas. It is assumed that there is no direct communication between the source and the destination. (This is reasonable when e.g., the destination is located far away from the source or there is a severe shadow fading between the source and the destination.) The communication between the source and the relay takes place in two phases as described in the following.

First–Hop Transmission: During the first phase, the source transmits its information and the relay listens to the transmitted signal. The received signal vector at the relay, denoted by $\boldsymbol{y}_1$, is given by

$$\boldsymbol{y}_1 = \boldsymbol{H}_1\boldsymbol{x}_1 + \boldsymbol{z}_1 \tag{1}$$

where $\boldsymbol{H}_1 \in \mathbb{C}^{[L\times M]}$ denotes the channel between the source and the relay, $\boldsymbol{x}_1 \in \mathbb{C}^{[M\times 1]}$ denotes the transmitted signal vector from the source and $\boldsymbol{z}_1 \in \mathbb{C}^{[L\times 1]}$ denotes the additive circularly symmetric Gaussian noise. The signal vector $\boldsymbol{x}_1$ is the output of the modulator α which is defined as

$$\alpha : \mathbb{W} \longmapsto \mathbb{C}^M$$
$$\boldsymbol{x}_1 = \alpha(w)$$

where $w \in \mathbb{W} \triangleq \{1, 2, 3, \ldots, 2^q\}$ denotes a message to be transmitted over the channel. Some particular choices for defining α are, for example, the 2^q-QAM and 2^q-PSK modulation schemes. We assume an average power constraint at the source, such that $\mathrm{tr}\mathbb{E}\{\boldsymbol{x}_1\boldsymbol{x}_1^\dagger\} \le P_1$.

Second–Hop Transmission: During the second phase, only the relay transmits and the source is silent. We assume that the relay uses a forwarding strategy given by the following deterministic function

$$f : \mathbb{C}^L \longmapsto \mathbb{C}^L$$
$$\boldsymbol{x}_2 = f(\boldsymbol{y}_1)$$

Since the function $f(\cdot)$ is arbitrary, our model includes linear as well as nonlinear mappings. We assume an average power constraint at the relay such that $\mathrm{tr}\mathbb{E}\{\boldsymbol{x}_2\boldsymbol{x}_2^\dagger\} \le P_2$. The received signal at the destination, denoted by $\boldsymbol{y}_2$, is then given by

$$\boldsymbol{y}_2 = \boldsymbol{H}_2\boldsymbol{x}_2 + \boldsymbol{z}_2 \tag{2}$$

where $\boldsymbol{H}_2 \in \mathbb{C}^{[N\times L]}$ denotes the channel between the relay and the destination, $\boldsymbol{x}_2 \in \mathbb{C}^{[L\times 1]}$ denotes the transmitted signal vector from the relay and $\boldsymbol{z}_2 \in \mathbb{C}^{[N\times 1]}$ denotes the additive circularly symmetric Gaussian noise. Finally, the destination, upon receiving $\boldsymbol{y}_2$, detects the transmitted message using the function (demodulator or detector) β defined as

$$\begin{aligned}\beta : \mathbb{C}^N &\longmapsto \mathbb{W}\\ \hat{w} &= \beta(\boldsymbol{y}_2)\end{aligned}$$

where $\hat{w} \in \mathbb{W}$ denotes the detected message at the destination.

Channel Statistics: We assume that the entries of the channel matrices $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are i.i.d. Rayleigh fading, distributed according to $\mathcal{CN}(0,1)$. The entries of the noise vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are assumed to be independent zero-mean circularly symmetric Gaussian noise. The covariance matrices of the noise vectors are given by $\boldsymbol{R}_{z_1z_1} = \mathbb{E}[\boldsymbol{z}_1\boldsymbol{z}_1^\dagger] = N_1\boldsymbol{I}_L$ and $\boldsymbol{R}_{z_2z_2} = \mathbb{E}[\boldsymbol{z}_2\boldsymbol{z}_2^\dagger] = N_2\boldsymbol{I}_N$, where $\boldsymbol{I}_N$ and $\boldsymbol{I}_M$ denote the identity matrices of size N and M, respectively. Additionally, we assume that the channels stay unchanged during the transmission of one block but they vary independently from one block to another.

Channel State Information (CSI): We assume that the source, the relay, and the destination know $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ perfectly. The CSI of backward channels at the relay and the destination can be obtained using training sequences and the CSI of the forward channels at the source and the relay can be obtained either using reciprocity of the links or feedback. When the channel matrices are constant or varying slowly, one can obtain accurate CSI at the nodes. Satellite MIMO link and wireless LAN are two practical examples in which this model is applicable.

2.1 Problem formulation

The goal is to minimize the *average message error probability*. Thus for a given message set $\mathbb{W}$, we need to find the triple (α^*, β^*, f^*) under the average power constraint such that

$$(\alpha^*, \beta^*, f^*) = \arg\min_{\alpha,\beta,f} \Pr\{\hat{w} \neq w\}. \tag{3}$$

We desire to find a *structured* solution to the optimization problem in (3). Imposing structure on a communication strategy results in loss of performance in general. On the other hand, a structured strategy however facilitates the design. We first utilize the channel knowledge to orthogonalize each hop using the SVD and then propose a nonlinear scheme that performs close to the theoretical bound.

3. Channel orthogonalization via SVD

In the following, we employ the singular value decomposition (SVD) to obtain an *equivalent* parallel channel for each hop. We then rewrite the optimization problem given by (3) for the equivalent channel.

Using the SVD, any channel realizations of $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ can be written as

$$\begin{aligned}\boldsymbol{H}_1 &= \boldsymbol{U}_1\boldsymbol{D}_1\boldsymbol{V}_1^\dagger\\ \boldsymbol{H}_2 &= \boldsymbol{U}_2\boldsymbol{D}_2\boldsymbol{V}_2^\dagger\end{aligned}$$

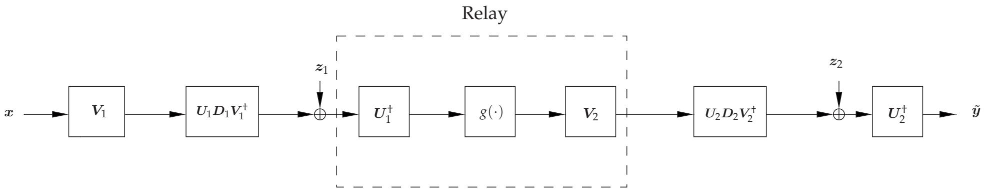


Fig. 2. Processing using the SVD of the channel matrices.

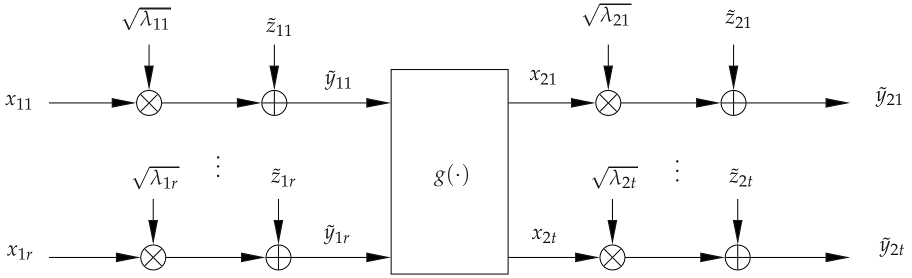


Fig. 3. Equivalent parallel channel.

where $\boldsymbol{U}_1 \in \mathbb{C}^{[L\times L]}$, $\boldsymbol{V}_1 \in \mathbb{C}^{[M\times M]}$, $\boldsymbol{U}_2 \in \mathbb{C}^{[N\times N]}$ and $\boldsymbol{V}_2 \in \mathbb{C}^{[L\times L]}$ are unitary matrices, and $\boldsymbol{D}_1 \in \mathbb{R}^{[L\times M]}$ and $\boldsymbol{D}_2 \in \mathbb{R}^{[N\times L]}$ are non-negative and diagonal matrices. Note that since $\boldsymbol{U}_1$, $\boldsymbol{V}_1$, $\boldsymbol{U}_2$ and $\boldsymbol{V}_2$ are invertible, linear operations of the form of $\boldsymbol{AG}$ or $\boldsymbol{GA}$ (where $\boldsymbol{G} \in \{\boldsymbol{U}_1, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{U}_2\}$ and $\boldsymbol{A}$ is an arbitrary matrix with an appropriate size) impose no loss of information. Thus we can preprocess the transmitted signal vectors from the source and the relay and postprocess the received signal vectors at the relay and the destination as illustrated in Fig. 2. Consequently, the received signal at the relay after the linear postprocessing is given by

$$\begin{aligned}\tilde{\boldsymbol{y}}_1 &= \boldsymbol{U}_1^\dagger \boldsymbol{y}_1\\ &= \boldsymbol{U}_1^\dagger \boldsymbol{H}_1 \boldsymbol{V}_1 \boldsymbol{x}_1 + \boldsymbol{U}_1^\dagger \boldsymbol{z}_1\\ &= \boldsymbol{U}_1^\dagger \boldsymbol{U}_1 \boldsymbol{D}_1 \boldsymbol{V}_1^\dagger \boldsymbol{V}_1 \boldsymbol{x}_1 + \boldsymbol{U}_1^\dagger \boldsymbol{z}_1\\ &= \boldsymbol{D}_1 \boldsymbol{x}_1 + \tilde{\boldsymbol{z}}_1\end{aligned}$$

where the last equality follows from the identities $\boldsymbol{U}_1^\dagger \boldsymbol{U}_1 = \boldsymbol{I}_L$ and $\boldsymbol{V}_1^\dagger \boldsymbol{V}_1 = \boldsymbol{I}_M$ and the definition $\tilde{\boldsymbol{z}}_1 = \boldsymbol{U}_1^\dagger \boldsymbol{z}_1$. The random vector $\tilde{\boldsymbol{z}}_1 \sim \mathcal{CN}(0, N_1 \boldsymbol{I}_L)$ since $\boldsymbol{U}_1$ is a unitary matrix. In a similar fashion, we can obtain

$$\tilde{\boldsymbol{y}}_2 = \boldsymbol{D}_2 \boldsymbol{x}_2 + \tilde{\boldsymbol{z}}_2$$

where $\tilde{\boldsymbol{z}}_2 \triangleq \boldsymbol{U}_2^\dagger \boldsymbol{z}_2 \sim \mathcal{CN}(0, N_2 \boldsymbol{I}_M)$. See also Fig. 2. Because $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices, we have

$$\begin{aligned}\tilde{y}_{1i} &= \sqrt{\lambda_{1i}} x_{1i} + \tilde{z}_{1i}, \quad i \in \{1, 2, \ldots, \min(M, L)\}\\ \tilde{y}_{2j} &= \sqrt{\lambda_{2j}} x_{2j} + \tilde{z}_{2j}, \quad j \in \{1, 2, \ldots, \min(L, N)\}\end{aligned}$$

where $\sqrt{\lambda_{1i}}$ is the ith entry on the main diagonal of $\boldsymbol{D}_1$ and $\sqrt{\lambda_{2j}}$ is the jth entry on the main diagonal of $\boldsymbol{D}_2$. The equivalent channel obtained by the SVD operation is shown in Fig. 3. The function $g(\cdot)$ in Fig. 3 denotes the forwarding strategy at the relay, defined as

$$g : \mathbb{C}^r \longmapsto \mathbb{C}^t$$
$$\boldsymbol{x}_2 = g(\tilde{\boldsymbol{y}}_1)$$

where $r = \min(M, L)$ and $t = \min(L, N)$. We consider both linear as well as nonlinear mappings. One can thus optimize the mapping according to

$$(\alpha^*, g^*(\tilde{\boldsymbol{y}}_1), \beta^*) = \underset{\left\{\alpha:\mathrm{tr}\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\dagger]\leq P_1\right\},\{g(\tilde{\boldsymbol{y}}_1):\mathrm{tr}\mathbb{E}[g(\tilde{\boldsymbol{y}}_1)g^\dagger(\tilde{\boldsymbol{y}}_1)]\leq P_2\},\beta}{\mathrm{argmin}} \Pr\{\hat{w} \neq w\}. \tag{4}$$

4. Transmission strategies and performance bounds

4.1 Lower bound on P_e

We next give a simple lower bound on the average message error probability, which we use as a benchmark to evaluate different transmission strategies in the sequel.

Lemma 1. *For the two-hop vector channel shown in Fig. 1, the average message error probability P_e is lower bounded by*

$$P_e \geq \max\{P_{e_1}, P_{e_2}\} \tag{5}$$

where P_{e_1} and P_{e_2} denote the average message error probability of the first- and the second hop, respectively.

Proof. Consider a two-hop channel where the first hop is noise-free and the second hop is identical to the original channel in Fig. 1. Denote the average error probability of this new channel by $\bar{P}_e$. It is easy to see that $P_e \geq \bar{P}_e = P_{e_2}$. In a similar manner we can obtain $P_e \geq \tilde{P}_e = P_{e_1}$, where $\tilde{P}_e$ denotes the error probability of a two-hop channel with identical first hop to that in Fig. 1 and a noise-free second hop. This yields (5). □

4.2 Linear relaying

One of the fundamental strategies in the literature is *linear* relaying, commonly known as amplify-and-forward (AF). Using AF in our setting, the relay function is given by

$$x_{2i} = g_i(\tilde{y}_{1i}) = \kappa_i \mu_i \tilde{y}_{1i}, \quad i \in \{1, ..., \min\{r, t\}\} \tag{6}$$

where $\mu_i = \sqrt{\frac{P_2}{\lambda_{1i}\mathbb{E}[x_{1i}x_{1i}^\dagger]+N_1}}$ is a power normalization factor and $0 \leq \kappa_i \leq 1$ is a power allocation factor where $\sum_{i=1}^t \kappa_i^2 = 1$. Note that the number of parallel channels that can be utilized is $\min\{r, t\}$, i.e., the minimum number of parallel streams of the first- and second hop. In (9), it is shown that the strategy given by (6) is optimal if the relay mapping is constrained to be linear. However as we show, AF is in general suboptimal for the underlying channel. The received signal-to-noise ratio (SNR) of the ith stream at the destination is given by

$$\gamma_i^{AF} = \frac{\kappa_i^2 \lambda_{1i} \lambda_{2i} P_{1i} P_2}{N_1 N_2 + \lambda_{1i} P_{1i} N_2 + \kappa_i^2 \lambda_{2i} P_2 N_1} \tag{7}$$

where $P_{1i} \triangleq \mathbb{E}[x_{1i}x_{1i}^\dagger]$. The fact that the received noise at the relay is forwarded to the destination is the main drawback of AF relaying.

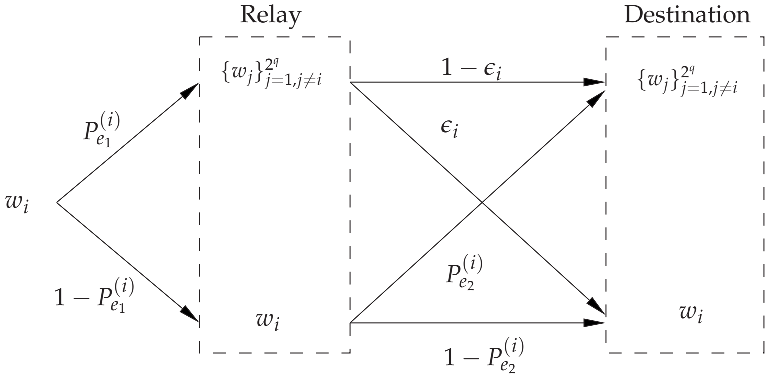


Fig. 4. Error probability transition in DF relaying.

Case $r=1$: In order to maximize γ_i, one should choose the strongest mode (the stream with largest singular value) with full power when $r=1$. Note that the use of weaker streams at the relay does not improve the performance of AF since all streams are transmitting the same signal, thus allocating all power to the strongest mode is the optimal solution. Therefore, the maximum possible achievable SNR for linear relaying when $r=1$, is given by

$$\gamma_{AF}^* = \frac{\lambda_{11}\lambda_{21}P_1P_2}{N_1N_2+\lambda_{11}P_1N_2+\lambda_{21}P_2N_1} \tag{8}$$

where λ_{11} and λ_{21} are the largest eigenvalues of the first- and second hop, respectively.

4.3 Relaying via Detect-and-Forward (DF)

Another approach for forwarding the received signals is to first detect the transmitted message and then re-modulate it. That is

$$x_{2i} = g_i(\tilde{\boldsymbol{y}}_1) = \kappa_i\alpha_{ri}(\hat{\tilde{w}}) = \kappa_i\alpha_{ri}(\beta_r(\tilde{\boldsymbol{y}}_1)) \tag{9}$$

where $\hat{\tilde{w}} = \beta_r(\tilde{\boldsymbol{y}}_1)$ is the detected message and β_r denotes the detector at the relay. The modulator for generating x_{2i} is denoted by α_{ri}. We also have $\mathrm{tr}\mathbb{E}[\boldsymbol{x}_2\boldsymbol{x}_2^\dagger] = P_2$.

The following proposition derives a simple upper bound on the average message error probability of DF relaying.

Lemma 2. *The average message error probability is upper bounded by*

$$P_e \leq P_{e_1} + P_{e_2} - \min_{1\leq i\leq 2^q} P_{e_1}^{(i)}P_{e_2}^{(i)} \tag{10}$$

where $P_{e_1}^{(i)}$ and $P_{e_1}^{(i)}$ respectively denote the ith message error probability of the first- and the second hop and P_{e_1} and P_{e_1} respectively are the average message error probabilities of the first- and the second hop.

Proof. Consider the transmission of w_i from the source. The relay either detects the transmitted message correctly or declares another message. This is illustrated in Fig. 4. Using Fig. 4, the *i*th message error probability can be bounded as

$$\begin{aligned} P_e^{(i)} &= (1-P_{e_1}^{(i)})P_{e_2}^{(i)} + P_{e_1}^{(i)}(1-\epsilon_i) \qquad (11)\\ &\leq P_{e_1}^{(i)} + P_{e_2}^{(i)} - P_{e_1}^{(i)}P_{e_2}^{(i)} \qquad (12) \end{aligned}$$

where ϵ_i denotes the detection probability of w_i at the destination when $\{w_j\}_{j=1,j\neq i}^{2^q}$ is transmitted from the relay, under the constraint that the source is transmitted w_i. The inequality in (11) follows from the fact that $0 \leq \epsilon_i \leq 1$. By taking average over all possible messages, we have

$$
\begin{aligned}
P_e &= \sum_{i=1}^{2^q} P_e^{(i)} p(w_i) && (13)\\
&\leq \sum_{i=1}^{2^q} \left(P_{e_1}^{(i)} + P_{e_2}^{(i)} - P_{e_1}^{(i)} P_{e_2}^{(i)} \right) p(w_i) && (14)\\
&= P_{e_1} + P_{e_2} - \sum_{i=1}^{2^q} P_{e_1}^{(i)} P_{e_2}^{(i)} p(w_i) && (15)\\
&\leq P_{e_1} + P_{e_2} - \left(\min_{1\leq i\leq 2^q} P_{e_1}^{(i)} P_{e_2}^{(i)} \right) \sum_{i=1}^{2^q} p(w_i) && (16)\\
&= P_{e_1} + P_{e_2} - \min_{1\leq i\leq 2^q} P_{e_1}^{(i)} P_{e_2}^{(i)} && (17)
\end{aligned}
$$

This completes the proof. □

Proposition 1. *DF relaying achieves the same performance as that of a single hop (i.e.,* $\max\{P_{e_1}, P_{e_2}\}$*) at high SNR when* $N \neq M$*.*

Proof. For given modulator and optimal demodulator, the error probability at the destination is upper bounded as

$$
\begin{aligned}
P_e^{DF} &\leq P_{e_1} + P_{e_2} = \frac{a_1}{\gamma_1^{NL}} + O\left(\frac{1}{\gamma_1^{NL+1}}\right) + \frac{a_2}{\gamma_2^{LM}} + O\left(\frac{1}{\gamma_2^{LM+1}}\right)\\
&= \begin{cases} \frac{a_1}{\gamma_1^{NL}} + O\left(\frac{1}{\gamma_1^{NL+1}}\right) & \text{if } N < M \\ \frac{a_2}{\gamma_2^{LM}} + O\left(\frac{1}{\gamma_2^{LM+1}}\right) & \text{if } M < N \end{cases} \qquad (18)
\end{aligned}
$$

where we used Lemma 2 and $\gamma_1 \triangleq \frac{P_1}{N_1}$, $\gamma_2 \triangleq \frac{P_2}{N_2}$, and a_1 and a_2 are two constants depending on the number of antennas and the modulation scheme.
We also have the following lower bound using Lemma 1

$$
P_e \geq \max\{P_{e_1}, P_{e_2}\} = \begin{cases} \frac{a_1}{\gamma_1^{NL}} + O\left(\frac{1}{\gamma_1^{NL+1}}\right) & \text{if } N < M \\ \frac{a_2}{\gamma_2^{LM}} + O\left(\frac{1}{\gamma_2^{LM+1}}\right) & \text{if } M < N \end{cases} \qquad (19)
$$

Comparing (18) and (19), we see that the upper bound and lower bound meet each other at high SNR. This therefore establishes the optimality of DF at high SNR. □

Proposition 2. *DF achieves the optimal diversity order* $d^* = \min\{NL, ML\}$*.*

Proof. From Lemma 1, we conclude that $d^* \leq \min\{NL, ML\}$. But, from Lemma 2 we know that $P_e^{DF} \leq P_{e_1} + P_{e_2}$. Thus the diversity order is bounded as $d_{DF} \geq \min\{NL, ML\}$. Therefore, DF achieves the optimal diversity order. □

In the following we comment further on the conventional DF and a propose a novel DF relaying scheme.

Conventional DF: One way to simplify the problem is to use the same modulator over all streams. That is $\alpha_{ri} = \alpha_r$ for all streams. By doing so, with a similar argument as that in the linear relaying case, the optimal power allocation would be to use all the power on the strongest mode.

Proposition 3. *Relaying using conventional DF (i.e., transmission using the strongest mode) is optimal at high SNR when $N > M$.*

Proof. The proof follows from the observation that using only the stream with the strongest mode of the second hop, one can obtain higher diversity gain compared to the first hop for any source modulator. Since $M < N$, we have

$$P_e^{DF} \leq \frac{a_1}{\gamma_1^{LM}} + O\left(\frac{1}{\gamma_1^{LM+1}}\right), \text{ and } P_e \geq \frac{a_1}{\gamma_1^{LM}} + O\left(\frac{1}{\gamma_1^{LM+1}}\right). \tag{20}$$

This completes the proof. □

Proposed DF: A more sophisticated approach at the relay is to use different modulators over distinct streams. In the following, we propose a structured method for obtaining different modulators based on a given modulator, say α_r. Let π denote a permutation operation on a given finite sequence. For example, if $a = (1, 2, 3, 4)$ the operation $\pi(a)$ produces a different ordering of the elements in the sequence, such as $\pi_1(a) = (4, 3, 1, 2)$. In the following let $\bar{\alpha}_r$ denote the list of letters produced by the modulator α_r, in the default order. Now we construct the ith modulator using $\bar{\alpha}_r$ as

$$\bar{\alpha}_{ri} = \kappa_i \pi_i(\bar{\alpha}_r) \tag{21}$$

where κ_i is a power allocation factor used at ith stream such that $\{\kappa_i\}$ meets the power constraint $\mathrm{tr}\mathbb{E}[\boldsymbol{x}_2\boldsymbol{x}_2^\dagger] \leq P_2$. Thus, the transmitted signal from the relay over the ith stream is given by

$$x_{2i} = g_i(\tilde{\boldsymbol{y}}_1) = \kappa_i \alpha_{ri}(\hat{\tilde{w}}) = \kappa_i \alpha_{ri}(\beta_r(\tilde{\boldsymbol{y}}_1)) \tag{22}$$

Here β_r denotes the detector used at the relay and the modulator α_{ri} is constructed using the ith permutation used over the ith stream, i.e., π_i. Now designing a relaying strategy specializes to finding the optimal permutations and the power allocation factors. That is

$$(\{\kappa_i^*\}_{i=1}^t, \{\pi_i^*\}_{i=1}^t) = \arg \min_{\{\kappa_i\}_{i=1}^t, \{\pi_i\}_{i=1}^t} \Pr\{\hat{w} \neq w\} \tag{23}$$

The proposed DF scheme includes conventional DF as a special case, by choosing $\kappa_i = 0$ for $i \neq 1$. Thus, the error probability achieved by the proposed DF scheme is upper bounded by that of conventional DF. The main advantage of the proposed scheme is that it enjoys a structured design based on a given modulator. From Proposition 3, one can conclude that this scheme does not bring any advantage at high SNR when $N > M$. However, in the following section we show that the proposed DF approach can attain considerable gain over conventional DF and linear relaying at moderate SNR's, that is, in an SNR regime where diversity gain is not a useful performance measure.

5. Numerical results and concluding remarks

In the following we present numerical results for the case when the source has only one single antenna and the relay and the destination have 10 antennas each. This scenario is of importance, for example in the uplink transmission in cellular networks, where the mobile node has only a single antenna. Under this constraint, the relay has only one incoming stream and multiple outgoing streams (see Fig. 3). Fig. 5 shows the average message error probability for three different relaying schemes; linear relaying, conventional DF relaying, the proposed DF relaying approach based on permutation mappings using two streams. We use 16-QAM as the modulator and an optimal ML detector at the relay and the destination. For the proposed scheme we use two streams in the second hop. The optimal permutation is obtained using exhaustive search. We also plotted a lower bound on the performance for any relaying scheme, using Lemma 1. Here we set $P_1 = P_2 = P$, $N_1 = N_2 = 1$. From Fig. 5, we see that linear relaying performs worst, and the proposed DF relaying scheme provides the best performance. Surprisingly, the performance of the proposed DF is very close to the lower bound.

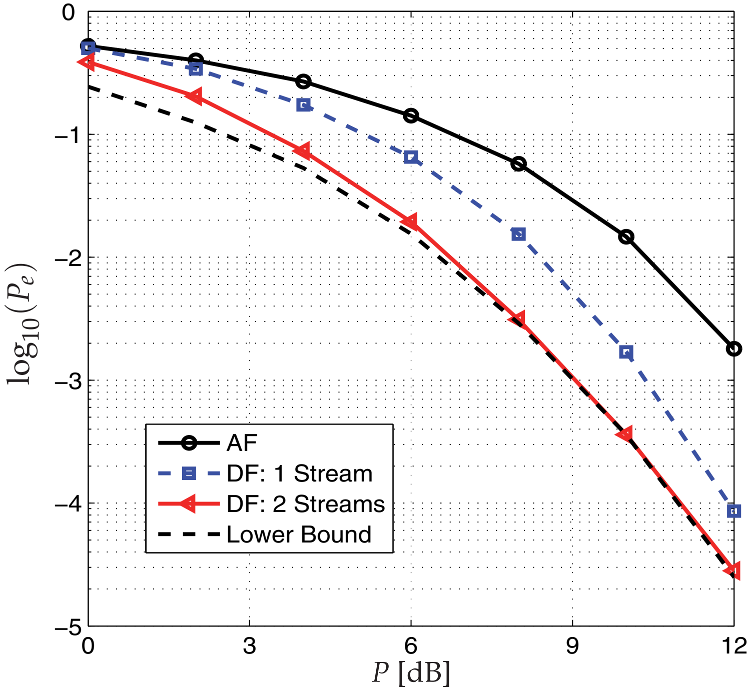


Fig. 5. Average message error probability (P_e) using 16-QAM modulation for different forwarding strategies (AF, conventional DF (i.e., one stream) and proposed DF (i.e., two streams with permuted modulations)). Here we set $P_1 = P_2 = P$, $N_1 = N_2 = 1$, number of antennas at the source is $N = 1$ and number of antennas at the relay and the destination are $L = M = 10$.

We can see from Fig. 5 that the performance of conventional DF approaches that of the proposed scheme at *high* SNR. This is in accordance with Proposition 3. However, we also see that the new scheme gives considerable gains in the low- and moderate SNR regime, and it achieves the optimal performance at lower SNR compared to conventional DF.

6. References

[1] E. C. van der Meulen, “Three-terminal communication channels,” *Adv. Appl. Probab.*, vol. 3, pp. 120-154, 1971.

[2] T. M. Cover and A. El Gamal, “Capacity theorems for the relay channel,” *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 572-584, Sep. 1979.

[3] J. N. Laneman, G. W. Wornell, and D. N. C. Tse, “Cooperative diversity in wireless networks: efficient protocols and outage behavior,” *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062-3080, Dec. 2004.

[4] E. Telatar, “Capacity of multi-antenna Gaussian channels,” *Technical Memorandum, Bell Laboratories (Published in European Transactions on Telecommunications*, Vol. 10, No.6, pp. 585-595, Nov/Dec 1999), 1995.

[5] D. Palomar and S. Barbarossa, “Designing MIMO communication systems: Constellation choice and linear transceiver design,” *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 3804-3818, Oct. 2005.

[6] B. Wang, J. Zhang, and A. Høst-Madsen, “On the capacity of MIMO relay channels,” *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 29-43, Jan. 2005.

[7] A. S. Behbahani, R. Merched, and A. M. Eltawil, “Optimization of a MIMO relay network” *IEEE Trans. on Signal Processing*, vol. 56, no. 10, Oct. 2008.

[8] N. Khajehnouri and A. H. Sayed, “Distributed MMSE relay strategies for wirless sensor networks” *IEEE Trans. on Signal Processing*, vol. 55, no. 7, July 2007.

[9] X. Tang and Y. Hua , “Optimal design of non-generative MIMO wireless relays” *IEEE Trans. on Wireless Communications*, vol. 6, no. 4, April 2007.

[10] M. N. Khormuji and E. G. Larsson, “Improving collaborative transmit diversity using constellation rearragment,” *In proceedings IEEE WCNC*, March 2007.

[11] M. N. Khormuji and E. G. Larsson, “Rate-optimized constellation rearrangement for the relay channel,” *IEEE Communication Letters*, vol. 12, no. 9, pp. 618-620, Sept. 2008.

11

Connectivity Support in Heterogeneous Wireless Networks

Anna Maria Vegni[1] and Roberto Cusani[2]
[1]*University of Roma Tre*
Department of Applied Electronics, Rome;
[2]*University of Roma "La Sapienza"*
Department of Information Engineering,
Electronics and Telecommunications, Rome;
Italy

1. Introduction

Recent advances in wireless technology and decreasing costs of portable devices strongly contributed to increase the popularity of mobile communications.

Wireless communication and device integration have lead to the so-called *nomadic computing* (or *mobile computing*) where portable devices (such as laptop and handheld computers) allow users to access Internet and data on their home or work computers from anywhere in the world. Multimedia services requirements nowadays encompass not only large bandwidths, but also *on-the-move* facilities. Future 4th generation wireless communications systems will provide seamless mobility support to access heterogeneous wired and wireless networks (Makhecha & Wandra, 2009), (Lin *et al.*, 2010).

Emerging and pre-existing wireless technologies exhibit different characteristics, access technologies, available services and network performances. For example GSM, UMTS, WLAN and WiMAX have different bandwidth (70 Mbps for WiMAX and 9.6 kbps for GSM), cell diameter (50 km in LoS for WiMAX and 100 m for WLAN), or handover latency (3 s for WLAN and 50 μs for WiMAX).

The increasing demand for services with high QoS requirements and novel mobility scenarios, like *on-the-move* business users, home and office networks, *on-the-move* entertainment, info-mobility etc., provide users to be connected to the Internet anytime and anywhere, as well as user services and connectivity be maintained, and kept alive. *Mobility management* in heterogeneous networks is the essential support for roaming nomadic devices switching from one access technology to another, at the same time maintaining seamless connectivity at high QoS services (*i.e.* video-streaming).

New emerging multimode mobile devices are equipped with multiple wireless network interface cards, providing *Vertical Handover* capability to autonomously select the best access network. The design of innovative handover mechanisms —sometimes called as *handoff*— between heterogeneous mobile devices (*e.g.* PDA, laptop, smart phones) and seamless integration of different integrated network (*e.g.* GSM, UMTS, HSDPA, GPS, WLAN, Bluetooth and so on) is an open research issue.

In this way, a mobile user can seamlessly switch between different networks, supporting the same services. This process must be performed to automatically adapt to change access networks and environments, without any user participation. In order to do this, cross layer design for multimedia communications is required. Mobile computing then becomes more feasible, *e.g.* a mobile user performing a videoconference using UMTS maintains this service even though the link breaks down, accessing into a WLAN network.

Vertical Handover (VHO) is a mechanism allowing heterogeneous connectivity by enabling switches from a serving network to a candidate network, whenever users or network requirements (*i.e.* power level, network congestions, or other QoS constraints) impose or suggest it. Notice that VHO allows switching from one access technology to another, thus offering additional functionalities with respect to classic horizontal handover where mobile nodes move from an access point to another without changing the serving access network (Balasubramaniam & Indulska, 2004), (McNair & Fang, 2004).

In this chapter we show how heterogeneous networks for next generation multimedia systems can cooperate in order to provide seamless mobility support to mobile users requiring high multimedia Quality-of-Service (QoS) constraints (Knightson *et al.*, 2005).

We describe the traditional techniques of Vertical Handover in heterogeneous wireless networks. Basically, in Section 2 we introduce the main characteristics of handover process and our effort is addressed on a first handover classification, which distinguishes between horizontal and vertical, hard and soft, upward and downward procedures, and more. Beyond several handover algorithms, in Subsection 2.1 we give an overview of current IEEE 802.21 standard for seamless connectivity in heterogeneous environments. In Section 3 we describe different decision metrics for handover mechanisms. Various metrics triggering handover decisions, including multi-parameters QoS, and mobile terminal location information, will be described in details in Subsection 3.1, and 3.2, respectively. Moreover, a hybrid approach which exploits both power measurements and location information will be presented in Subsection 3.3. Finally conclusions are drawn in Section 4.

2. Vertical handover procedures overview

New-generation wireless networks adopt a heterogeneous broadband technology model aiming to guarantee seamless connectivity to mobile users, anytime and anywhere. Different network characteristics are expected for different multimedia applications, each of them requiring a specific QoS level. Ubiquitous access through a single network technology could not always guarantee seamless connectivity, due to geographical coverage limitations, so that the cooperation of different access networks represents an important feature for heterogeneous environments.

A general definition of handover assumes it as the process by which a mobile terminal keeps its connection active when migrating from the coverage of one network Access Point (AP) to another. Basically, different types of handovers can occur in wireless overlay networks. Network switching can be performed not only to maintain user connectivity but also to keep high QoS. There are some decision handover parameters based on QoS, available resources, channel quality or preference consumer.

In GSM, handover decision is based on the perception of channel quality, reflected by the received signal strength and the availability of resources in neighbour cells. The Base Station (BS) usually measures the quality of the radio link channels used by Mobile Nodes (MNs) in its service area. Measures are periodically updated so that degradations in signal strength going below a prescribed threshold can be detected and handover toward another radio channel or cell can be initiated.

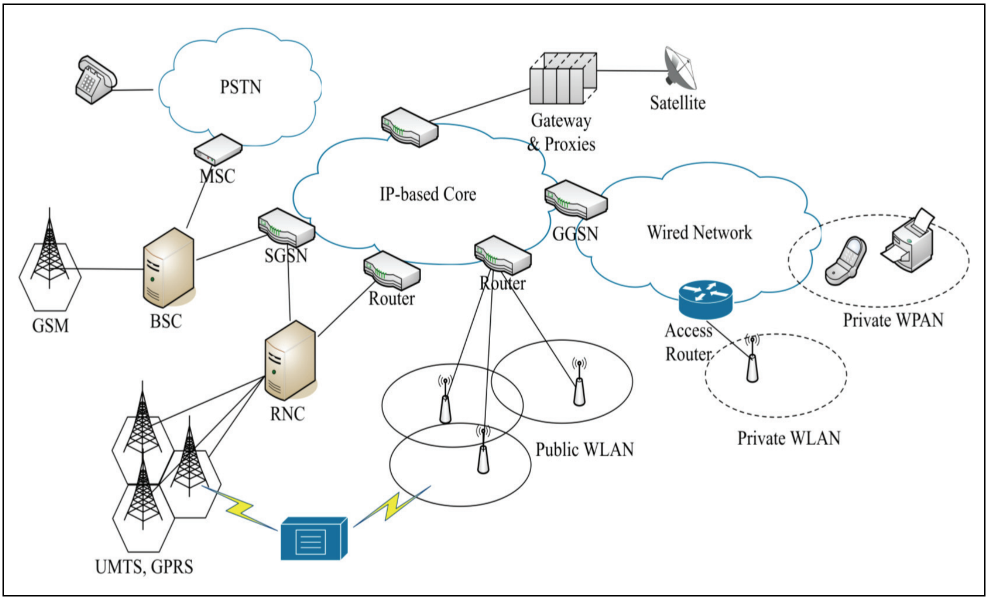


Fig. 1. Heterogeneous networks scenario

Horizontal handover (HHO) occurs between the APs of the same network technology, while *vertical handover* (VHO) occurs between APs belonging to different networks. Several kind of VHO can be envisaged, as described as follows. According to Figure 1, *upward* vertical handover is a handover to a wireless overlay with a larger cell size and generally lower bandwidth per unit area. It makes a mobile device disconnect from a network providing faster but smaller coverage (*e.g.* WLAN) to a new network providing slower but broader coverage.

Viceversa, a mobile device performing a *downward* VHO disconnects from a cell providing broader coverage to one providing limited coverage but higher access speed. In this case, a link layer trigger can inform the mobile device that it is now under the coverage of a new network (*e.g.* WLAN) and the mobile node may wish to execute the handover.

Downward VHOs may be *anticipated* or *unanticipated*, such that a mobile device may already be under the coverage of the new network but may prefer to postpone the handover based on requirements of the applications running on the mobile node. Handover is then performed later, being already aware of the coverage status of the new network.

A main issue is to decide if or when to start the handover, and who performs it. Handover policies are based on different metrics for handover decision. Traditional solutions simply consider RSSI (Received Signal Strength Indication) and channel availability. More sophisticated handover policies also consider: (*i*) Quality-of-Service, as different types of services require various combinations of reliability, latency, and data rate; (*ii*) costs, *i.e.* different networks may employ different billing strategies; (*iii*) network conditions like traffic, available bandwidth, network latency, and congestion; (*iv*) system performance, such as channel propagation characteristics, path loss, inter-channel interference, Signal-to-Noise ratio and Bit Error Rate; (*v*) mobile terminal conditions like battery power and dynamic factors such as speed, moving pattern, moving histories, and location information. In the latter case, if the battery level of a MN is low, the handover commutes toward a network that guarantees lower power consumption. In the case when the user requires a guaranteed QoS level for her applications, handover switches to a network meeting such requirements.

A more detailed description of handover decision metrics will be given in Section 3.

2.1 IEEE 802.21 media-independent handover

The IEEE 802.21 group is developing standards to enable handover and interoperability between heterogeneous network types, including both 802 and non 802 networks.

The standard provides quick handovers of data sessions across heterogeneous networks with small switching delays and minimized latency. The handover in heterogeneous networks could become more flexible and appropriate with this standard, through the use of innovative IEEE 802.21 mobile devices. The standard considers both wired and wireless technologies such as 802.3, 802.11, 802.16, 3GPP2, and 3GPP.

The analysis of IEEE 802.21 standard aims to understand the scope of this protocol. Seamless handover of data sessions is the main target, based on Media Independent Handover (MIH) functional model. IEEE 802.21 specification classifies the function that enhances handovers across heterogeneous media. The MIH protocol entity is to every extent a new protocol layer located between the Network Layer (Layer 3) and the interface-specific lower layers (MAC and PHY in the case of IEEE interfaces, RRC and LAC in the case of 3GPP or 3GPP2 interfaces, respectively).

The main entities of IEEE 802.21 are (Gupta *et al.*, 2006):

1. The *Media Independent Information Service* (MIIS) that includes policies and directives from the Home Network (HN). The mobile terminal refers to the HN policies when performing handover decisions;
2. The *Service Access Points* (SAPs), exchanging service primitives between the MIH layer and its adjacent layers and functional planes;
3. A *Decision Engine* (DE) within the MIH instance, residing in the mobile terminal, which identifies the best available access technology to support the current connectivity. The DE is a state machine that selects a preferred link based on available interfaces, policies, QoS and security parameter mapping;
4. A *Transport Mechanism* to facilitate the communication between the mobile terminal MIH and the Information Service (IS) instance to access in the network.

The MIH function at the mobile terminal is continuously supplied with information regarding the network conditions, measured to perform the access into one available heterogeneous network. The MIH function receives the information through dedicated interfaces by exchanging messages with the IS entity positioned in the HN. Generally, the MIHF defines three main services to perform handovers between heterogeneous networks, such as (*i*) the Media Independent Event Service (MIES), (*ii*) the Media Independent Command Service (MICS), and (*iii*) the Media Independent Information Service (MIIS).

MIES provides event reporting, event filtering and event classification corresponding to dynamic changes in link characteristics, link quality and link status. It acts all the instances to make event detection and notify, still maintaining the actual link connection to the MN.

Some of these events employed are “Link Up”, “Link Down”, “Link Detect”, “Link Parameter Reports” and “Link Going Down”, (Gupta, *et al.* 2006).

MICS uses the MIHF primitives to send commands from higher layers to lower ones. It determines the status of the connected links, performing mobile and connectivity decisions of the higher layers to the lower ones. Then, MIIS is a mechanism to discover available neighboring network information in order to facilitate handover process. It assures a set of information entities, both static and dynamic. In the first case, there are the names and services providers of the MN’s current network, while the dynamic information include link layer parameters such as channel information, MAC addresses, security information and other higher layer service information.

Three main handover schemes have been developed within IEEE 802.21 standardization process, defined as follows:

1. *Serving Network-Initiated and Candidate Network-Canceled Handover*: the Service Access Network (SAN) sends messages about information request to the IS in order to know if a handover mechanism can be initiated. The Candidate Network (CN) could not be an available resource, because of link quality level or network traffic status;
2. *Serving Network-Initiated and MN-Canceled Handover*: after sending information request messages to the IS, the MN could be not available to perform handover, because of MN movement or user intervention or low battery that let the MN renouncing handover mechanism. In this case, the handover could be canceled directly by interaction between SAN and CN;
3. *MN-Initiated and MN-Canceled Handover*: in this case, the MN communicates only with the SAN, which sends messages of Resources-Request to the CN. The MN could be no more available to perform handover, (*e.g.* movement or time out or user intervention). The MN sends HandoverCancel message to SAN which is asked to interrupt the handover mechanism. Figure 2 shows the flowchart of this approach.

3. Techniques for connectivity support in heterogeneous networks

Vertical Handover preserves user connectivity *on–the–move* (Pollini, 1996). It is applied when network switching is expected in order to (*i*) preserve host connectivity, (*ii*) optimize QoS as perceived by the end user, and (*iii*) limit the number of unnecessary vertical handover occurrences.

Different VHO schemes can be classified on the basis of the criteria and parameters adopted in the handover initialization phase. The following list collects the main metrics whose monitoring can drive handover decisions:

- *Received Signal Strength* (RSS)-based VHO algorithms are largely used in cellular networks (*i.e.* 2G and 3G networks). The handover process is initiated on the basis of a decreasing level of measured RSS (Ayyappan & Dananjayan, 2008); (Inzerilli & Vegni, 2008);
- *Signal-to-Noise and Interference ratio* (SINR)-based VHO algorithms are typically used in UMTS networks. SINR factor directly impacts achievable goodput in a wireless access network. The handover is driven by a reduction of measured SINR below a fixed threshold (Yang *et al.*, 2007); (Vegni *et al.*, 2009);
- *Multi-parameter QoS*-based VHO algorithms: this approach is based on the overall quality assessment for the available networks obtained balancing various parameters — subjective and objective quality metrics— (Vegni *et al.*, 2007);

- *Location*-based VHO algorithms estimate network QoS levels on the basis of the MN's location relatively to the serving access point (Kibria *et al.*, 2005); (Inzerilli *et al.*, 2008); (Inzerilli *et al.*, 2010).

The above-mentioned VHO algorithms are analyzed here in the following.

The RSS-based VHO is the traditional technique for connectivity switching in heterogeneous networks. It is driven on power level measurements, such as when the measured RSS coming from the SN drops below a predefined threshold. The RSS of the monitored set of CNs is evaluated, and a vertical handover will be executed towards the best—most appropriate—candidate network. This approach represents the primitive and simplest handover mechanism, which however does not aim to optimize communication performance but only focuses on maintaining seamless connectivity (Ayyappan & Dananjayan, 2008). Moreover, since RSS value suffers from severe fluctuations due to the effects of shadowing and fading channels, filtering techniques (*e.g.* exponential smoothing average (Inzerilli & Vegni, 2008)) should be considered to estimate the trend of RSS signal.

On the other hand, the SINR-based approach compares received power with noise and interference levels in order to obtain a more accurate performance assessment. SINR factor represents a valid handover decision metric, as it directly affects the maximum data rate compatible with a given Bit Error Rate (BER). SINR-based VHO approach is more suitable to meet QoS requirements, since a reduction of SINR factor produces a reduction of data rate and QoS level (Yang *et al.*, 2007). Both RSS and SINR-based schemes are reactive approaches, whose aim is to compensate for performance degradation.

The multi-parameter QoS-based VHO scheme in (Vegni *et al.*, 2007) represents a proactive approach performing regular assessment of the QoS level offered by the current SN, as well as by other CNs. In general, a multi-parameter QoS-based VHO technique is well suited for multimedia applications like real-time video streaming.

In location-based VHO solutions, the knowledge of MN's location information is exploited to assess the quality of the bidirectional link between SN and MN (Inzerilli *et al.*, 2008). Moreover, the estimation of MN's position can drive the initiation of a reactive handover mechanism. Information about MN's position can be determined in several ways (Kibria *et al.*, 2005), including Time of Arrival, Direction of Arrival, RSS, as well as A-GPS (Assisted Global Positioning System) techniques.

Notice that in general each time a vertical handover is initiated the traffic overhead increases. The limitation of handover occurrences is an issue specially when unwanted and unnecessary vertical handovers are executed. This represents the case of a mobile node moving back and forth between the two neighbouring wireless networks—or in general around a corner that involves three or more wireless networks—. This aspect in known in literature as *ping-pong* effect (Kim *et al.*, 2007).

Repeated vertical handover attempts lead to frequent location and registration updates (with network resource consumption), frequent connectivity interruptions, as well as serious affections to MN's QoS (*i.e.*, decreasing battery life). Frequent handovers lead the user to experience many unpleasant transients of service interruption.

Techniques to prevent unnecessary and unwanted handovers have been proposed (Kim *et al.*, 2007); (Inzerilli & Vegni, 2008); (Inzerilli *et al.*, 2008). A hysteresis cycle or a hard limitation in maximum handover frequency can mitigate this phenomenon.

The above descriptions have shown the main vertical handover approaches, which are based on single metrics (*i.e.*, RSS, SINR, QoS, and location). Still, many handover techniques are based on the combination of two or more metrics, which generate most effective VHO

decisions, and avoid unnecessary and unwanted vertical handover occurrences (Vegni *et al.*, 2009). These approaches are called as *hybrid* (or combined) vertical handovers.
The following Subsection 3.1 and 3.2 offer a more detailed description of QoS- and location-based vertical handover algorithms, respectively.

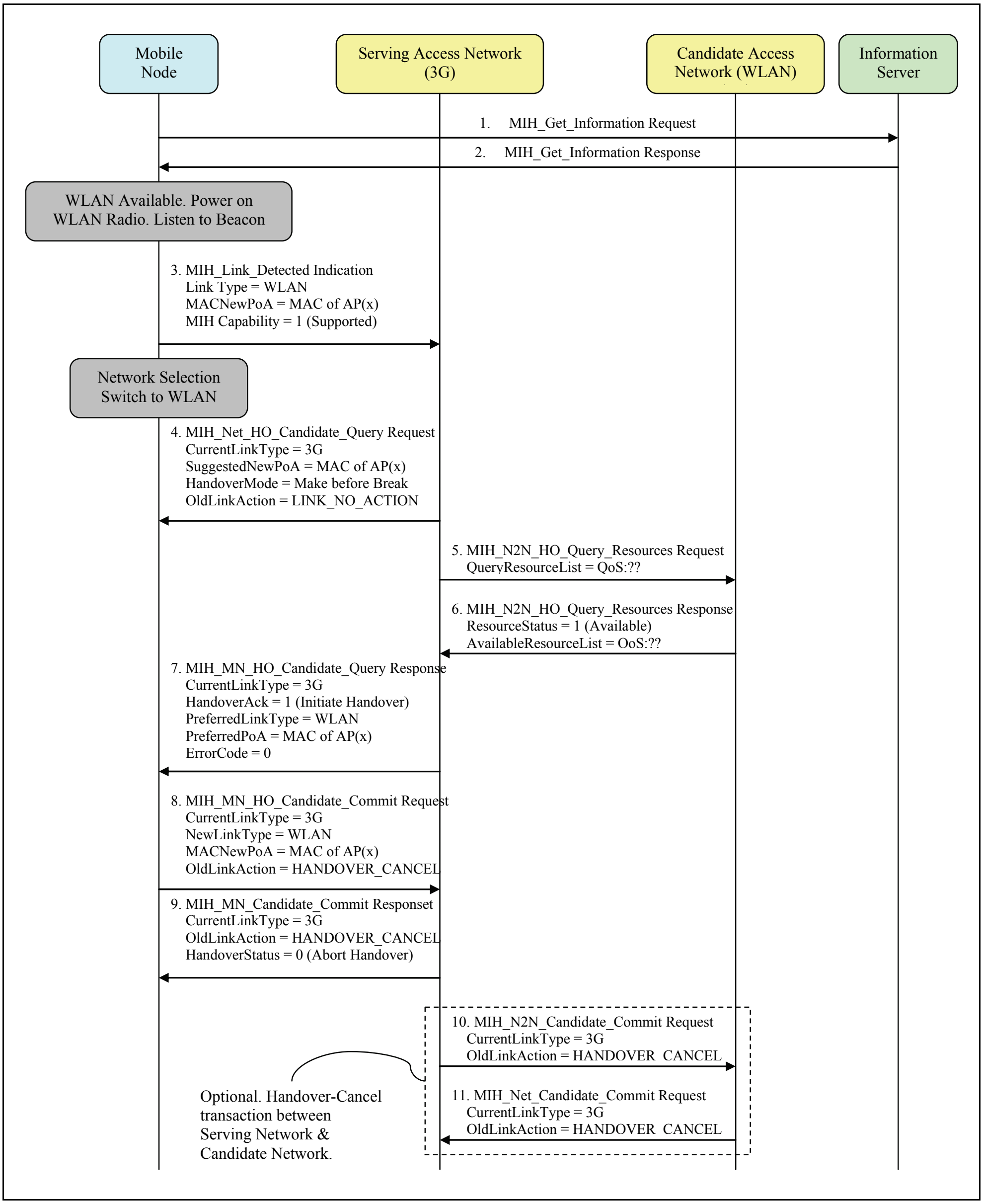

Mobile Node
Serving Access Network (3G)
Candidate Access Network (WLAN)
Information Server
1. MIH_Get_Information Request
2. MIH_Get_Information Response
WLAN Available. Power on WLAN Radio. Listen to Beacon
3. MIH_Link_Detected Indication
Link Type = WLAN
MACNewPoA = MAC of AP(x)
MIH Capability = 1 (Supported)
Network Selection Switch to WLAN
4. MIH_Net_HO_Candidate_Query Request
CurrentLinkType = 3G
SuggestedNewPoA = MAC of AP(x)
HandoverMode = Make before Break
OldLinkAction = LINK_NO_ACTION
5. MIH_N2N_HO_Query_Resources Request
QueryResourceList = QoS:??
6. MIH_N2N_HO_Query_Resources Response
ResourceStatus = 1 (Available)
AvailableResourceList = OoS:??
7. MIH_MN_HO_Candidate_Query Response
CurrentLinkType = 3G
HandoverAck = 1 (Initiate Handover)
PreferredLinkType = WLAN
PreferredPoA = MAC of AP(x)
ErrorCode = 0
8. MIH_MN_HO_Candidate_Commit Request
CurrentLinkType = 3G
NewLinkType = WLAN
MACNewPoA = MAC of AP(x)
OldLinkAction = HANDOVER_CANCEL
9. MIH_MN_Candidate_Commit Responset
CurrentLinkType = 3G
OldLinkAction = HANDOVER_CANCEL
HandoverStatus = 0 (Abort Handover)
10. MIH_N2N_Candidate_Commit Request
CurrentLinkType = 3G
OldLinkAction = HANDOVER CANCEL
11. MIH_Net_Candidate_Commit Request
CurrentLinkType = 3G
OldLinkAction = HANDOVER CANCEL
Optional. Handover-Cancel transaction between Serving Network & Candidate Network.

Fig. 2. MN-initiated and MN-cancelled handover

3.1 QoS-based vertical handover

As defined in IEEE 802.21 standard, QoS-based HO decision is based on current and expected network conditions, according to the application QoS requirements (Golmie *et al.*, 2006). Current network conditions are measured using network performance parameters from various layers, *e.g.* signal strength from layer 1, packet loss from layer 2, throughput and delay from layer 2+, etc.

According to the ITU-T Y.1540, the applications QoS requirements are defined by:

- Packet Transfer Delay (PTD), maximum end-to-end tolerated delay [s];
- Packet Delay Variation (PTV), *i.e.* jitter: maximum packet jitter [s];
- Packet Loss Ratio (PLR): maximum tolerated packet loss;
- Throughput: required data rate of successful packets [bit/s].

QoS-based Decision Engine (QDE) is a main network entity that implements a QoS-based handover for assigned application QOS requirements (Golmie *et al.*, 2006). QDE is a MIH user that considers application QoS requirements and network performance measurements provided by the MIH. The MIH function exchanges information between network entities and the QDE, including technology, protocol types and network measurements. Network performance conditions, such as instantaneous measurements for current conditions, are evaluated from past observations and previous connections, or as default estimates.

QDE entity is located as a remote entity, as part of the MN, or the AP/BS. For our scope, we consider it as a network entity, as illustrated in Figure 3. No limitation occurs if QDE function is distributed over remote entities of each network. In this way, the MN receives Video Quality Metrics VQMs from QDEs of neighbouring candidate networks.

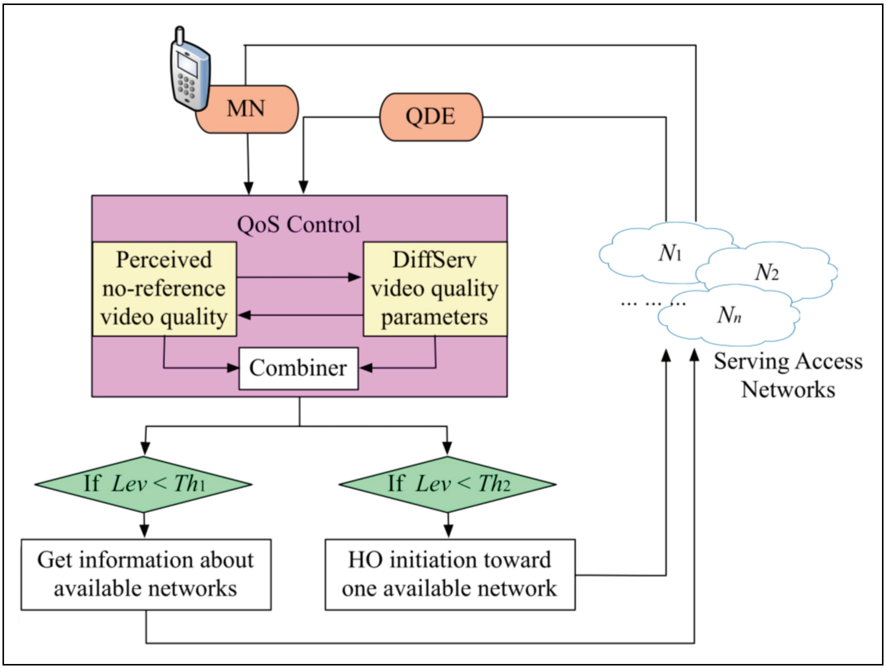

MN
QDE
QoS Control
Perceived no-reference video quality
DiffServ video quality parameters
Combiner
N_1
N_2
… … …
N_n
Serving Access Networks
If $Lev < Th_1$
If $Lev < Th_2$
Get information about available networks
HO initiation toward one available network

Fig. 3. Proposed network architecture for QoS-based handover

During the last decade several techniques to assess the quality of multimedia services without performing a subjective test have been investigated, leading to objective quality metrics that approximate the MOS (Mean Opinion Score). Such metrics are usually classified as *full-reference*, *reduced reference* and *no-reference* metrics. They differ in the degree of knowledge about the original multimedia flow used in the quality assessment. Full reference methods evaluate the difference between the original signal and the received one and are, thus, rarely employed in real time assessment of video quality. On the other hand, reduced reference metrics require a channel for sending side information concerning some characteristics of the original signal, while No-Reference (NR) metrics do not require any knowledge of the original. On the other hand, the perceived video quality highly depends on the end-user, according to her *a priori* knowledge of the topic, level of attention while looking at the video, and assigned task.

In general a QoS-based VHO technique focuses on the maximization/enhancement of QoS level experienced by the mobile user. The connectivity is switched from the serving network to a selected candidate network, which provides high QoS. In (Vegni *et al.*, 2007) an innovative NR-Video Quality Metric (NR-VQM) has been assumed. This metric incorporates both spatial and temporal resolution reduction, packet losses, latency, and delay jitter, as well as indicators for the evaluation of (*i*) blocking, blurring and ringing introduced by current video coders, and (*ii*) jerkiness and other effects produced by packet losses. Jerkiness is evaluated by a neural network feed with estimated dynamics of objects composing the scene, trained by means of subjective tests.

The impact of packet losses on perceived quality is based on the analysis of the inter-frame correlation measured at the receiver side. Presence of error concealment algorithms employed by the receiver is also taken into account. The packet losses degradation assessment presents best performances for long sequences with slow motion.

Since the NR algorithm directly processes the rendered video, no information about the kind of errors, delays and latencies affecting the links is required. In addition, the NR techniques easily account for continuous increase of computing power of both mobile and wired terminals, and allow a wide spread of error concealment techniques to increase the perceived quality.

Subjective test evinced good agreement between NR metric and MOS, regardless of intrinsic video characteristic and spatial-temporal resolution.

In the scheme presented in (Vegni *et al.*, 2007) the NR-QoS is combined with a packet classification originally designed for DiffServ (DS) protocol, allowing different QoS grades to be mapped into different classes of aggregated traffic flows, (Shin *et al.*, 2001). It is an adaptive packet forwarding mechanism for DS networks. It allows mapping mechanism of video packets onto different DS levels based on Relative Priority Index (RPI), that is the relative preference per each packet in terms of loss and delay. The packets are classified, conditioned, and remarked to certain network DS levels by considering the traffic profile based on the Service Level Agreement (SLA) and the current network status.

The proposed QoS-based handoff scheme in (Vegni *et al.*, 2007) is well suited for real-time applications in IEEE 802.21 scenarios. The overall vertical handover process is organized in five phases, such as (*i*) measurement phase, (*ii*) QoS prioritization, (*iii*) initialization, (*iv*) candidate networks scanning, and (*v*) VQM conversion. Figure 4 depicts the pseudo-code of the QoS-based vertical handover algorithm. Basically, the multi-parameter QoS-based handover algorithm considers three inputs, such as (*i*) the RSS samples, (*ii*) the mobile node distance and (*iii*) the current QoS level (*i.e. Lev*). As output it returns the number of vertical handovers executed by the mobile node (*i.e.* n_{VHO}). In the following we shall describe the overall mechanism in more details.

Given a video application, the QoS mapping process is accomplished by considering the relative prioritized packets to maximize end-to-end video quality. In the *measurement phase*, each *T* seconds, the MN gets samples of RSS and position, as well as monitors *Lev* parameter for the video stream received from the current serving network. QoS monitoring is performed on the basis of the NR metrics[1]. In our scope, NR technique addresses on audio and video flows evaluations to the receiver side, although it could be also tuned and optimized by means of a full-reference metric applied to some low rate probe signals.

In the *QoS prioritization phase*, the probability to perform a handover is evaluated (*i.e.* P_{VHO}). By opportunistically weighting the QoS-*Lev* parameter the handover probability will be mainly driven by QoS factors. The received video-streaming quality is monitored according to a subjective evaluation. On the basis of user preferences, two appropriate QoS thresholds are defined, called as Th_1 and Th_2, with $Th_1 > Th_2$.

```
Input : $\{RSS; d; Lev\}$
Output : $n_{VHO}$
while $T > 0$ do
    Measurement phase in SN
    if $Lev < Th_1$ then
        First alarm to QDE
        QoS priorization phase
            Calculation of $\Pr_{VHO}(x)$
        end
        Candidate network scanning phase
            List of target networks
            if $Lev < Th_2$ then
                Handoff initiation phase
                    VQM Conversion phase
                        Selection of a target network
                        $n_{VHO} \leftarrow 1$ (VHO executed)
                    end
                end
            end
        end
    else
        $n_{VHO} \leftarrow 0$ (no VHO executed)
    end
end
```

Fig. 4. Multi-parameter QoS-based vertical handover pseudo-code

[1] NR method presents some basic indicators for temporal and spatial analysis: block distortion is evaluated by applying first a coarse temporal analysis for each frame, to extract blocks potentially affected by artefacts produced by lost packets.

Traffic congestions, transmission errors, lost packets or delay can keep QoS level lower than a first threshold, *i.e.* $Lev < Th_1$. If *Lev* keeps on decreasing, the *handoff initiation* phase can be required by the MN whenever $Lev < Th_2$.

The MN alerts to change the serving network and sends this alarm message to a closer QDE. So, the *candidate networks scanning* phase occurs on the basis of VQM parameters, such as throughput, link packet error rate, Packet Loss Probability (PLP), supported number of Class of Service (CoS), etc. All these parameters are sent inside the information message "LINK QoS PARAMETER LIST".

Based on the statistics computed on previous NR-QoS reports produced by the served MNs, when the QDE communicates with a MN, it can operate a *conversion of VQM parameters* for each network in NR parameters. The MN evaluates which candidate network is appropriate for its video application. As an instance, let us suppose that a MN is in a WLAN area. When it realizes a QoS reduction, it sends a first alarm to the QDE, which will start a candidate network scanning process in order to select a target network providing a QoS enhancement to the MN (*i.e.* $Lev_1 > Lev$). A set of target networks to hand over is selected, and the best network is chosen on the basis of MN preferences and handover policies. Finally, the handover is performed according to the IEEE 802.21 message exchange in the scheme of Figure 2.

Finally, in order to determine the probability to perform a vertical handover (*i.e.* Pr(*VHO*)), we shall provide the following assumptions:

1. A mobile node is locating at position $P = (x, y)$, in the middle of two active networks (*i.e.* N_i, $i = 1,2$);
2. The averaged received signal levels over N_1 and N_2 radio links are assumed as lognormal distributions, respectively $\overline{r_{N_1}(P)}$ and $\overline{r_{N_2}(P)}$, with mean signal levels μ_{N_1} and μ_{N_2}, and the shadowing standard deviations $\sigma^2_{N_1}$ and $\sigma^2_{N_2}$;
3. The distance d_{N_i} from the MN's position to the reference BS of network N_i can be assumed as a stationary random process with mean value d and variance $c^2\sigma^2_{t_{dn}}$, where c the speed of light and $\sigma^2_{t_{dn}}$ is the standard deviation of the signal delay measurement [2];
4. On the basis of each single parameter (*i.e.* RSSI, distance, and QoS) different thresholds are assumed, called as *R*, *D* and *Q* for RSSI, distance and quality criterions, respectively. Each threshold is typical for a single access network (*i.e.* R_W and R_U are the RSS thresholds for WLAN and UMTS, respectively).

Let us suppose to perform a handover from UMTS to WLAN. The handover decision occurs when both (*i*) the RSS measurement on WLAN is higher than R_W (*i.e.* $\overline{r_W(P)} \geq R_W$), (*ii*) the distance from MN to WLAN AP is lower than D_W (*i.e.* $d_W \leq D_W$), and (*iii*) the QoS-*Lev* in WLAN is upper than Q_W (*i.e.* $q_W \geq Q_W$) [3]. Thus, the probability to initiate the handover from UMTS to WLAN in the position *P*, is

$$\Pr_{U\to W}(P) = P\left\{\left[\overline{r_W(P)} \geq R_W\right]\right\} \cdot P\left\{\left[d_W \leq D_W\right]\right\} \cdot P\left\{\left[q_W \geq Th_2\right]\right\}. \tag{1}$$

On the other hand, the handover decision from WLAN to UMTS is taken only when it is really necessary, such as when (*i*) the RSS measurement on WLAN is lower than R_W (*i.e.*

[2] Basically, the delay measurement of the signal between the MN and the BS is characterized by two terms, (*i*) the real delay and (*ii*) the measurement noise t_{dn}. It is assumed to be a stationary zero-mean random process with normal distribution.

[3] Notice that due to chip WLAN monetary cost the handover decision does not take account to the RSS, the distance and the QoS criteria on UMTS network.

$\overline{r_W(P)} \le R_W$), (*ii*) the RSS measurement on UMTS is higher than R_U (*i.e.* $\overline{r_U(P)} \ge R_U$), (*iii*) the distance from MN to UMTS BS is lower than D_U (*i.e.* $d_U \le D_U$), and (*iv*) the QoS-*Lev* in UMTS is upper than Q_U (*i.e.* $q_U \ge Q_U$). Thus, the probability to initiate the handover from WLAN to UMTS in the position *P*, is

$$\Pr_{W\to U}(P) = P\left\{\left[\overline{r_W(P)} \le R_W\right]\right\} \cdot P\left\{\left[\overline{r_U(P)} \ge R_U\right]\right\} \cdot P\{[d_U \le D_U]\} \cdot P\{[q_U \ge Th_2]\}. \tag{2}$$

3.1.1 Analytical model

In this subsection we introduce the analytical model behind the QoS-based VHO technique described in (Vegni *et al.*, 2007). Particularly, we shall define two main network parameters for handover decision from a serving network to a candidate network, such as (*i*) the average time delay and (*ii*) the average packet rate. Based on these parameters, the handoff mechanism shall be performed only if it is necessary to maintain the connection on.

We recall the *average time delay* [s] for the *k*-th network from the Pollaczeck-Kinchin formula, which considers the average time delay as the sum of average time delay for the service and waiting one, such as

$$\tau_k = \frac{1}{\mu_k} \cdot \left[\frac{1 + \rho_k \left(1 + Cb^2\right)}{2\left(1 - \rho_k\right)} \right]. \tag{3}$$

From (3) we consider the average time delay for a single packet sent from a N_1 to N_2 (*i.e.* $T_{N_1 \to N_2}$ [s]) such as

$$T_{N_1 \to N_2} = \sum_{k=1}^{N} \tau_k \gamma_{N_1 \to N_2}^{(k)}, \tag{4}$$

where $\gamma_{N_1 \to N_2}^{(k)}$ is the probability that packets are sent from N_1 to N_2, on the *k*-th link with capacity C_k [bit/s].

The *average packet rate* represents how many packets are sent from N_1 to N_2, *i.e.* $r_{N_1 \to N_2}$ [packets/s]. Considering all the available networks (*i.e.* N_1, N_2, …, N_n), the total *average packet rate* Λ_{tot} [packets/s] is

$$\Lambda_{tot} = \sum_{i=1}^{N} \sum_{j=1}^{N} r_{N_i \to N_j}, \tag{5}$$

and the total mean time delay ΔT_{Mean} [s] is

$$\Delta T_{Mean} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} T_{N_i \to N_j} \cdot r_{N_i \to N_j}}{\sum_{i=1}^{N} \sum_{j=1}^{N} r_{N_i \to N_j}}. \tag{6}$$

In the case of handover occurrence from N_i to N_j, the mobile user moves from N_i to N_j with a probability $\beta_{i \to j}$. So, the probability ν_j that an user moves from her own serving network is:

$$\nu_j = \sum_{i \neq j} \beta_{i \to j} = 1 - \beta_{i \to i}, \tag{7}$$

where $\beta_{i \to i}$ represents the probability a user stays in her serving network. In this way, we can find the average packet rate from N_i to N_j during handover, as

$$r^{(HO)}_{N_i \to N_j} = \sum_m r^{(HO)}_{N_j \to N_m} \beta_{j \to m} + \beta_{i \to j} \sum_h r^{(HO)}_{N_i \to N_h}. \tag{8}$$

So, the total packet rate $\Lambda^{(HO)}_{tot}$ [packets/s] will be:

$$\begin{aligned}\Lambda^{(HO)}_{tot} &= \sum_{i=1}^{N}\sum_{j=1}^{N} r^{(HO)}_{N_i \to N_j} = \sum_i \sum_j \sum_m r^{(HO)}_{N_i \to N_m} \beta_{i \to m} + \sum_i \sum_j \beta_{i \to h} \sum_h r^{(HO)}_{N_i \to N_h} = \\ &= \Lambda_{tot} + \sum_i \sum_j \beta_{i \to h} \sum_h r^{(HO)}_{N_i \to N_h}.\end{aligned} \tag{9}$$

Let us assume ρ_j [packets/s] as the average rate of packets sent to N_j. By replacing (7), the expression of Λ^{HO}_{tot} becomes

$$\Lambda^{HO}_{tot} = \Lambda_{tot} + \sum_j \nu_j \rho_j. \tag{10}$$

If we consider an uniform handover probability (*i.e.* $\nu_j = \nu$) then $\Lambda^{(HO)}_{tot}$ becomes

$$\Lambda^{(HO)}_{tot} = \Lambda_{tot}\left(1+\nu\right). \tag{11}$$

Finally, let χ be the ratio between the average time delay in case and in absence of handover, $\tau^{(HO)}$ and τ respectively

$$\begin{aligned}\chi = \frac{\tau^{(HO)}}{\tau} &= \frac{\dfrac{1}{\mu_k}\left[\dfrac{1+\theta^{(HO)}\left(1+Cb^2\right)}{2\left(1-\theta^{(HO)}\right)}\right]}{\dfrac{1}{\mu_k}\left[\dfrac{1+\theta\left(1+Cb^2\right)}{2\left(1-\theta\right)}\right]} = \frac{1+\theta^{(HO)}\left(1+Cb^2\right)}{1+\theta\left(1+Cb^2\right)} \cdot \frac{\left(1-\theta\right)}{\left(1-\theta^{(HO)}\right)} = \\ &= \frac{1+\theta\left(1+Cb^2\right)\left(1+\nu\right)}{1+\theta\left(1+Cb^2\right)} \cdot \frac{\left(1-\theta\right)}{1-\theta\left(1+\nu\right)}.\end{aligned} \tag{12}$$

where $\theta^{(HO)}$ [Bit/s] is the throughput experienced by a mobile user during handover.

3.2 Location-based vertical handover

In this subsection we shall introduce a location-based vertical handover approach (Inzerilli *et al.*, 2008) which aims at the twofold goal of (*i*) maximizing the goodput and (*ii*) limiting the *ping-pong* effect. The potentialities of using location information for VHO decisions, especially in the initiation process is proven by experimental results obtained through computer simulation. Leveraging on such results, in this subsection we shall introduce only the handover initiation phase since it represents the core of our location-based VHO technique. A detailed description of the proposed algorithm is in (Inzerilli *et al.*, 2008).

The mobile node's location information is used to initiate handovers, that is, when the distance of the MN from the centre of the cell of the candidate network towards which a handover is attempted possesses an estimated goodput, *i.e.* GP_{CN}, significantly greater than the goodput of the current serving network, *i.e.* GP_{SN}. The *handover initiation* is then followed by a more accurate estimate (*handover assessment*) which actually enables or prevent handover execution (Inzerilli *et al.*, 2008).

In the *handover initiation* phase, the algorithm evaluates the goodput experienced by a MN in a wireless cell. The goodput depends on the bandwidth allocated to the mobile for the requested services and the channel quality. When un-elastic traffic is conveyed (*e.g.* real-time flows over UDP) the goodput is given by:

$$GP = BW \cdot \left(1 - P_{out}\right), \tag{13}$$

where BW [Bit/s] is the bandwidth allocated to the mobile node and P_{out} is the service outage probability. When elastic traffic is conveyed (typically when TCP is used), throughput tends to decrease with increasing values of P_{out}. BW is a function of the nominal capacity, of the MAC algorithm which is used in a specific technology and sometimes of the experienced P_{out}. We consider the maximum value of BW, *i.e.* $BW_{\max}$ which is obtained in the case of a single MN in the cell and with a null P_{out}[4].

P_{out} is a function of various parameters. In UMTS network it can be calculated theoretically, using the following formula:

$$P_{out}^{UMTS} = \Pr\left\{ \frac{E_{b,Tx}^{UMTS}}{I_0 + \left(\gamma\sigma_N^2\right)_{UMTS}} \cdot A_d^{-1}\left(r^{UMTS}\right) \le \mu^{UMTS} \right\}, \tag{14}$$

where $E_{b,Tx}^{UMTS}$ is the bit energy in the received signal, μ and γ are parameters dependent on the signal and interference statistics, σ_N^2 is the receiver noise power, A_d (r^{UMTS}) is the signal attenuation factor dependent on the MN's distance r^{UMTS} from the centre of the cell, and I_0 is the inter and intra-cell interference power. The service outage probability for a WLAN network P_{out}^{WLAN} can be calculated theoretically in a similar fashion using the following formula:

$$P_{out}^{WLAN} = \Pr\left\{ \frac{E_{b,Tx}^{WLAN}}{\left(\gamma\sigma_N^2\right)_{WLAN}} \cdot A_d^{-1}\left(r^{WLAN}\right) \le \mu^{WLAN} \right\}. \tag{15}$$

We define as the radius of a wireless cell R_{cell} the distance from the cell centre beyond which the signal-to-noise ratio or the signal-to-interference ratio falls below the minimum acceptable value (*i.e.* μ). R_{cell} can be obtained resolving the above equations or empirically, through measurement on the network. As an alternative, typical value for well-known technologies can be used, *e.g.* $R_{cell}^{WIFI} \approx 120$ m for IEEE 802.11a outdoor, and 100 m ≤ R_{cell}^{UMTS} < 1 km for a UMTS micro-cell.

[4] In an IEEE 802.11a link, the maximum theoretical BW_{WLAN} is equal to 23 Mbps (out of a nominal capacity of 54 Mbps), although it decreases rapidly with the number of users because of the contention-based MAC. In HSDPA network, the maximum BW_{UMTS} is equal to 14.4 Mbps, which decreases rapidly with P_{out}.

Since the path loss $A_d(r)$ is approximately proportional to r^γ, the SNR(r) can be written as

$$\mathrm{SNR}(r) = \mu\left[\left(\frac{R_{cell}}{r}\right)^{\gamma} + \delta A_d\right]. \tag{16}$$

Maximum *GP* in a WLAN and UMTS cell can be calculated with the following approximated formulas, respectively

$$\begin{cases} GP_{\max}^{UMTS} = BW_{\max}^{UMTS} \cdot \Pr\left\{\left(\dfrac{R_{cell}^{UMTS}}{r^{UMTS}}\right)^{\gamma} + \delta A_d < 1\right\} \\ GP_{\max}^{WLAN} = BW_{\max}^{WLAN} \cdot \Pr\left\{\left(\dfrac{R_{cell}^{WLAN}}{r^{WLAN}}\right)^{\gamma} + \delta A_d < 1\right\} \end{cases} \tag{17}$$

which will be regarded as zero out of cells.

Handover initiation will be performed when the estimated goodput of the new network is greater than the current one. Namely, in the case of vertical handover from WLAN to UMTS, the following equations applies:

$$GP_{\max}^{UMTS} < GP_{\max}^{WLAN}. \tag{18}$$

It is worth noticing that when handover executions are taken too frequently, the quality as perceived by the end user can degrade significantly in addition to wasting battery charge.

3.2.1 Simulation results

In this section we report on network performance of the Location-based Vertical Handover algorithm (also called as LB-VHO). Particularly, we investigate the Cumulative Received Bits (CRB [Bits]), and the number of vertical handovers performed by the user moving in the grid, obtained using our event-driven simulator. Details of the simulator can be found in (Vegni, 2010).

We modelled movements of a MN over a grid of 400 x 400 square zones, each with an edge of 5 m, where 3 UMTS cells and 20 IEEE 802.11b cells are located. Typical data rate values have been considered for UMTS and WLAN. The location of each wireless cell has been generated uniformly at random, as well as the the MN's path.

Table 1, shows the statistics on the CRB collected for $S = 20$ randomly generated scenarios, each of them differs from the other in terms of the UMTS/WLAN cell location and the path of the MN on the grid. Performance have been compared to a traditional Power-based Vertical Handover (PB-VHO), which uses power measurements in order to initiate VHOs instead of mobile location information (Inzerilli & Vegni 2008).

For each approach LB and PB three parameters are reported related to the CRB, *i.e.* the mean value, the standard deviation and the dispersion index, defined as the ratio of the standard deviation over the mean value. The three value for LB and PB are reported versus different values of the waiting time parameter T_{wait}[5].

[5] Notice that if the MN moves at 1 m/s, a 10 s waiting time results to 10 m walked.

The LB approach brings about a reduction of CRB between 6.5% for a null waiting time and 20% for waiting time equal to 60 s. It follows that the waiting time constraint is not suitable for LB approach in order to reduce the number of vertical handovers while keeping a limited reduction of CRB.

Table 2 shows results of the number of VHO experienced with the LB and PB approach, still in terms of the mean value, standard deviation and dispersion index for various waiting time values. It can be noticed that the number of vertical handover with LB is on average significantly smaller, *i.e.* ranging in [9.65, 3.70] than that experienced with PB approach, *i.e.* ranging in [9.15, 329.85]. This remarks that the PB approach requires a constraint on handover frequency limitations, while this approach is counterproductive with LB.

Waiting Time [s]	LB Mean [Gb]	LB Stand. Dev [Gb].	LB Disp. Index	PB Mean [Gb].	PB Stand. Dev. [Gb].	PB Disp. Index
0	5.82	2.38	40.91%	6.23	2.30	36.90 %
60	4.59	2.34	50.88%	5.76	2.14	37.13 %

Table 1. Statistics on the CRB for LB and PB approach

Waiting Time [s]	LB Mean [Gb]	LB Stand. Dev. [Gb]	LB Disp. Index	PB Mean [Gb]	PB Stand. Dev. [Gb]	PB Disp. Index
0	9.65	2.00	20.73	329.85	794.50	240.87
10	7.25	1.15	15.93	30.20	46.36	153.51
20	5.85	2.31	39.48	19.90	22.54	113.26
30	5.15	1.15	22.42	14.10	16.29	115.53
40	4.35	1.15	26.54	11.80	12.49	105.85
50	4.20	2.00	47.62	9.80	10.58	107.99
60	3.70	1.15	31.21	9.15	7.57	82.75

Table 2. Statistics on the Number of VHO for LB and PB approach

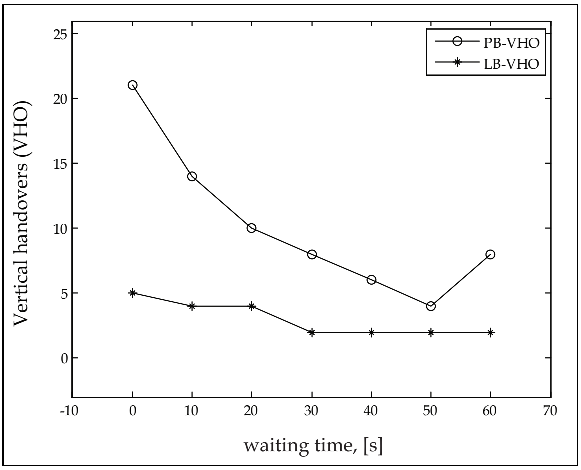


Fig. 5. Number of vertical handover occurrences for PB and LB VHO algorithm

In Figure 5, the mean values of vertical handovers for LB and PB vs. the waiting time constraint are depicted. This shows even more clearly how the LB approach, providing a more accurate assessment for handover initiation, limits handover initiations, resulting in about a little performance gain. In contrast, PB approach is unstable even for high values of waiting time, as it can be noticed from the fact that the PB curve is not monotone.

Finally, in Figure 6 (*a*) and (*b*) are reported the dynamics of the CRB over the mobile node steps during the simulation (a step is performed every 5 seconds) for a null waiting time and a waiting time of 60 s, respectively. The instability of PB approach when no waiting time constraint is applied is clearly shown in Figure 6 (*a*).

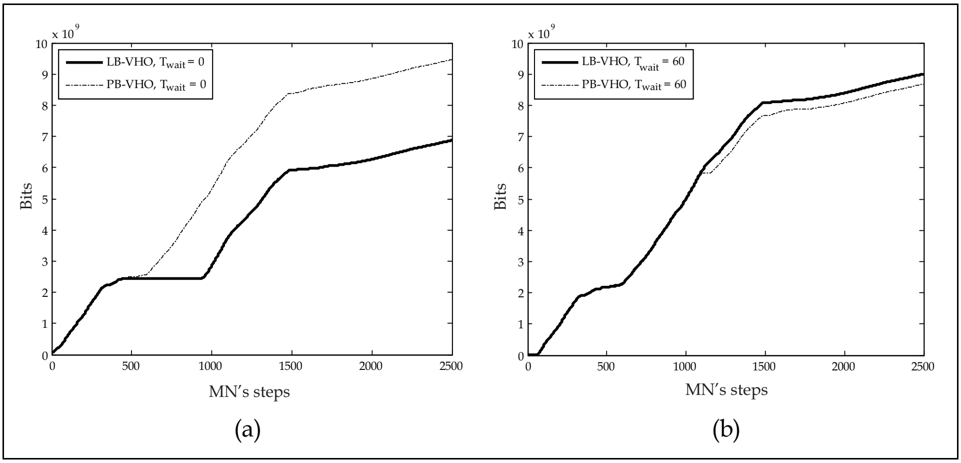


Fig. 6. CRB during a simulated scenario with PB and LB-VHO approaches, for (*a*) T_{wait} = 0 s, (*b*) T_{wait} = 60 s

3.3 Hybrid vertical handover technique

In this section we complete the overview of the main vertical handover techniques in heterogeneous wireless networks, by introducing a hybrid scheme for connectivity support[6]. Different wireless networks exhibit quite different data rate, data integrity, transmission range, and transport delay. As a consequence, direct comparison between different wireless links offering connectivity to a MN is not straightforward. In many cases VHO requires a preliminary definition of performance metrics for all the visited networks which allows to compare the Quality-of-Service offered by each of them and to decide for the best.

VHO decisions can rely on wireless channel state, network layer characteristics and application requirements. Various parameters can be taken into account, for example: type of the application (*e.g.* conversational, streaming, interactive, background), minimum bandwidth and maximum delay, bit error rate, transmitting power, current battery status of the MN, as well as user's preferences.

In this section we present a mobile-controlled reactive Hybrid VHO scheme —called as HVHO— where handover decisions are taken on the basis of an integrated approach using three components: (*i*) power map building, (*ii*) power-based (PB) VHO, and (*iii*) enhanced

[6] An extended version of this technique is described in (Inzerilli *et al.*, 2010).

location-based (ELB) VHO. The HVHO technique is suitable for dual-mode mobile terminals provided with UMTS and WLAN network interface cards, exploiting RSS measurements, MN's location information, and goodput estimation as discussed in Section 3. The overall procedure is mobile-driven, soft and includes measures to limit the *ping-pong* effect in handover decisions. The flowchart of HVHO is depicted in Figure 7.

Basically, the HVHO approach proceeds in two phases:

1. In the initial learning phase when the visited environment is unknown, the RSS based approach is used, *i.e.* hereafter referred to as Power-Based (PB) mode. In the meanwhile, the MN continuously monitors the strength of the signals received from the SN, as well as from the other candidate networks. By combining RSS samples with location data provided by the networks or some auxiliary navigation aids, like GPS, the MN builds a path losses map for each discovered network in the visited environment;
2. At the end of this phase the MN enters the ELB-VHO mode and it can exploit the path losses map to take handover decision using its current location.

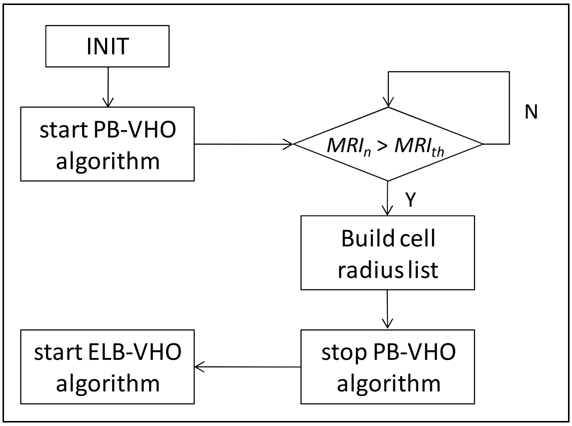


Fig. 7. Flowchart of hybrid vertical handover algorithm

In the initial learning phase, the new environment is scanned in order to detect the UMTS and WLAN access networks eventually present and, then to build a *path loss map* for each of them. The path losses associated to the UMTS base stations in the monitored set and to the access points of the WLAN network are estimated by taking the difference between the nominal transmitted power and the short term average of the received signal strength. Averaging is required in order to smooth fast fluctuations produced by multipaths, and can be performed by means of a mean filter applied to the RSS time series multiplied by a sliding temporal window (Inzerilli & Vegni, 2008).

Let n be the discrete time index and p_n be the power measure at time t_n. The moving average estimate P_N of the received power on a sliding window of length K is

$$P_N = \frac{1}{K}\sum_{i=N-K+1}^{n} p_i, \quad N \geq K. \tag{19}$$

Though averaging over the last K samples, it allows reducing the impact of instantaneous power fluctuations in power detection and reduce the power error estimation. On the other hand, as the MN is assumed to be moving, the length of moving windows depends on the actual MN speed. However, moving average filters are prone to outlayers. A more robust estimate can therefore be computed by replacing the linear mean filter with the (non-linear) median filter.

Let us suppose a mobile node is moving in an area approximated with a lattice of *M* x *M* square zones, each with a width L_{zone} [m]. While moving on the lattice, the MN calculates the power received for each visited zone Z_j. Let $P_n{}^{Zj}$ denote the average of the power samples collected inside Z_j and associates it with the planar coordinates of the centre of the zone (x_j, y_j). Namely, power level calculated at time n for the zone (x_j, y_j) is given by

$$P_n^{Z_j} = P_n\left(x_i, y_j\right) = \frac{1}{\left\|Z(j)\right\|}\sum_{i \in Z(j)} p_i, \tag{20}$$

where *Z* (*j*) is the set of the power samples p_i collected in the last visited zone *j* up to time *n*, and $\left\|Z(j)\right\|$ is the cardinality of the set *Z* (*j*).

Equation (20) provides a criterion to assess received power from both the UMTS and WLAN networks on which handover decisions of the PB approach are based. In addition, it allows assessing the power $P(x_j, y_j)$ for each Zj zone which can be stored in the terminal and populate a power map for the visited area. Once each zone of the lattice were visited at least once, the power map would be completed. However, it is possible that the complete visit of all the zones of the map can take long, and perhaps never occurs, especially if the number of zones M^2 is big. As a consequence, in order to accelerate power map building we can use polynomial interpolation to assign a power value to zones which has not been visited yet.

Namely, let us assume that zone Z_j has not been assigned a power value yet. Moreover, let Z_{j1} and Z_{j2} be the nearest zones and aligned to Z_j (as depicted in the examples of Figure 8), with a power value assigned. We can use linear interpolation to assess the power value P (x_j, y_j) of Z_j as follows. When the zone *j*-th is between zone j_1 and j_2 (Figure 8(*a*) and (*b*)), assessed power value of $P(x_j, y_j)$ of zone *j*-th is given by:

$$P\left(x_i, y_j\right) = \frac{D_j^{j_2}}{D_j^{j_1} + D_j^{j_2}} P\left(x_{j_1}, y_{j_1}\right) + \frac{D_j^{j_1}}{D_j^{j_1} + D_j^{j_2}} P\left(x_{j_2}, y_{j_2}\right), \text{ with } D_j^{j_1} < D_j^{j_2}, \tag{21}$$

where

$$D_j^{j_1} = \sqrt{\left(x_j - x_{j_1}\right)^2 + \left(y_j - y_{j_1}\right)^2},\ D_j^{j_2} = \sqrt{\left(x_j - x_{j_2}\right)^2 + \left(y_j - y_{j_2}\right)^2} \tag{22}$$

are the Euclidean distances between the centers of the zones $Z_{j_1} = \left(x_j, y_j\right)$ and $Z_{j_2} = \left(x_j, y_j\right)$ with $Z_j = \left(x_j, y_j\right)$, respectively. Conversely, when Z_j is not between Z_{j_1} and Z_{j_2}, as depicted in Figure 7 (*c*) and (*d*), the assessed power value P (x_j, y_j) of the zone *j*-th is given by:

$$P\left(x_j, y_j\right) = \left|\frac{D_j^{j_2}}{D_j^{j_2} - D_j^{j_1}} P\left(x_{j_1}, y_{j_1}\right) - \frac{D_j^{j_2}}{D_j^{j_2} - D_j^{j_1}} P\left(x_{j_2}, y_{j_2}\right)\right|. \tag{23}$$

It is worth highlighting that linear interpolation through (22) and (23) brings about errors in the power map. In addition, the exploitation of (22) and (23) starting from the power values of all visited zones does not guarantee completion of the power map. In general, a sufficient

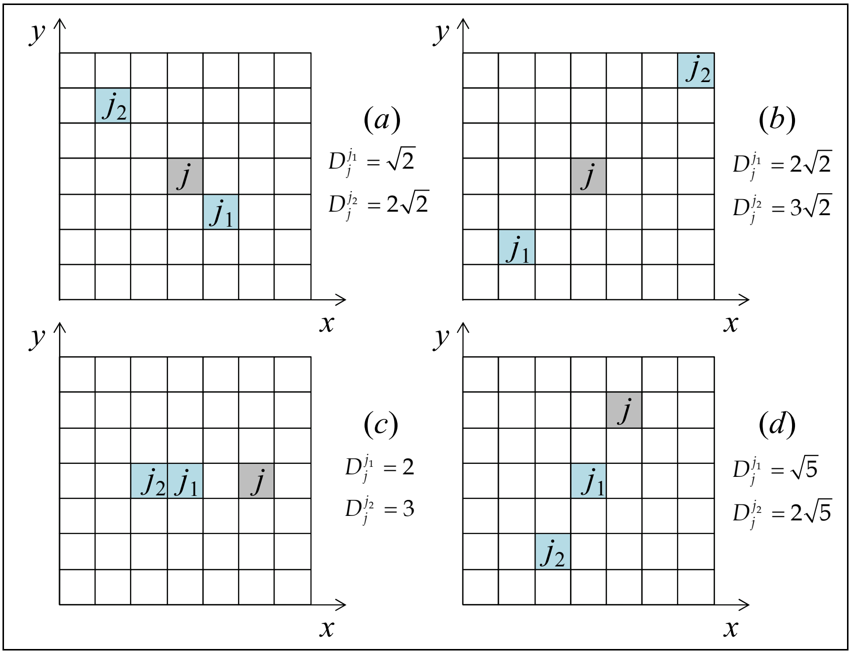


Fig. 8. Power Map built according to the displacement of zones

number of visited zones has to be achieved prior completion of the power map is possible. Such number is also dependent on the actual path of the MN in the lattice.

Let us introduce a coefficient to denote the degree of reliability of the power map at time *n*. Let VZ_n be the set of visited zones up to time *n*. We introduce the Map Reliability Index MRI_n at time *n* as follows:

$$MRI_n = \frac{\|VZ_n\|}{M^2}, \tag{24}$$

where M^2 is the total number of zones in the square lattice. We fix empirically a threshold value MRI_{TH} for the index in (24) beyond which the knowledge of the visited environment is regarded acceptable. Only when this threshold value is reached MRI_{TH}, polynomial interpolation with (22) and (23) is started. As in (Wang *et al.*, 2001), a lookup table of power profiles in each visited area is stored in the MN's database.

Basically, each visited zone size depends on the rate of change of the received power signals. For example, L_{zone} = 50 m is typical for a macro-cell with a slow average power variation, and L_{zone} = 10 m for a microcell with a fast signal change. Then, each zone size has pre-measured signal means and standard deviations for the serving cell and the neighbouring cells.

Figure 9 shows how a visited zone is built. The MN is in position $P_1 = (x_1, y_1)$, while $P_2 = (x_c, y_c)$ is the centre of a WLAN/UMTS cell. We can evaluate the angle α between the line from P_1 and P_2 and the horizontal plane, as:

$$\alpha = \arcsin\left(\frac{y - y_c}{\overline{P_1 P_2}}\right), \tag{25}$$

where $\overline{P_1P_2}$ is the distance from P_1 and P_2 obtained according to the Euclidean formula. The angle α is adopted to get the power attenuation, as we assume that the WLAN/UMTS cell radius $r_{\mathrm{WLAN/UMTS}}$ strictly depends on a factor $\gamma(\alpha)$, that modifies the cell radius value, as:

$$r_{WLAN/UMTS}(\alpha) = \gamma(\alpha) \cdot r_{WLAN/UMTS} \, . \tag{26}$$

The factor $\gamma(\alpha)$ is expressed as:

$$\gamma(\alpha) = 1 + 0.8 \cdot \sin\left(\frac{\alpha - \varphi_k^{WLAN/UMTS}}{2}\right), \tag{27}$$

where φ_k represents the k-th WLAN/UMTS cell down-tilt value, such as

$$\varphi_k^{WLAN/UMTS} = k \cdot \frac{360}{N_{WLAN/UMTS}}, \tag{28}$$

which depends on the number of WLAN/UMTS cells $N_{WLAN/UMTS}$, (*i.e.* 10 and 3 WLAN access points and UMTS base stations, respectively). So, the factor $\gamma(\alpha)$ is in the range [0.2, 1.8], and $r_{WLAN/UMTS}$ (α) will be decreased or increased of 80% of $r_{WLAN/UMTS}$.

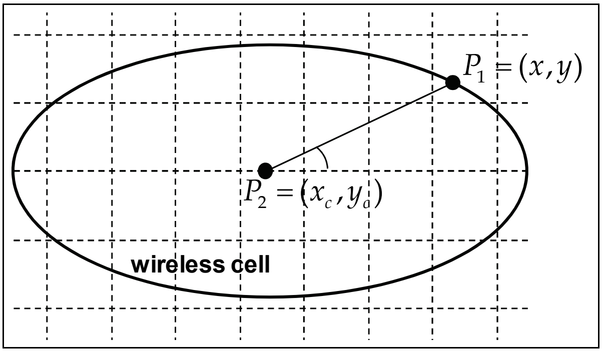


Fig. 9. Trigonometric approach for path loss map building in a (circular) wireless cell environment

3.3.1 Power-based approach for hybrid vertical handover

The Power-based VHO approach is exploited by the HVHO technique during the power maps building phase. Particularly, from mobile switch-off up to the completion of the power maps of both the UMTS and WLAN networks, the mobile node uses the PB-VHO approach to guarantee seamless connectivity (Inzerilli & Vegni, 2008). It performs handover using power measurements only, and does not take account of location information.

With the PB-VHO scheme the MN selects a network access, either UMTS or WLAN, and keeps it till the received power from the current network drops below the receiver sensitivity. Hence, the other network is scanned in order to verify if a handover to the other network can be done. Namely, if the power from the other network exceeds the receiver sensitivity, a handover to the new network is executed. In case power from both networks is below the minimum sensitivity, power scanning in both networks is continued repeatedly till one of the two networks exhibit a power value above its sensitivity threshold.

Power scanning frequency is limited in order to preserve battery charge as well as to prevent the ping-pong effect. When the mobile switches on it attempts selecting the WLAN network interface. Namely, if the measured power from the WLAN network interface card is above the value of MN WLAN receiver sensitivity, then the WLAN connectivity is available and the WLAN access is selected. Otherwise, if the measured power from the UMTS network interface card is above the value of MN UMTS receiver sensitivity, UMTS connectivity is available and the UMTS access is selected. When both checks fail, the mobile node waits a waiting time pause before re-trying the WLAN network scanning again.

3.3.2 Enhanced location-based approach for hybrid vertical handover

In the location-based approach presented in Subsection 3.2 we have assumed WLAN/UMTS circular cells. When the mobile terminal accesses a power map of the visited area representing the WLAN coverage area, as well as the UMTS coverage area, it is possible to derive a more accurate assessment of the *GP* in the UMTS and WLAN links, respectively. In this subsection we introduce a modified version of the location-based approach as described in Subsection 3.2, that is suitable for non-circular wireless cells. In particular, we consider a variable radius for each cell *k*-th, such as the cell radius becomes a function of the angle α_k between the horizontal axe and the axe connecting the border of the cell with its centre. This approach is then called as Enhanced Location-based (ELB) VHO, for non-circular wireless cells. The ELB-VHO approach is exploited in order to obtain a characterization of the wireless cell geometry coming from the power map building phase, as depicted in Figure 10.

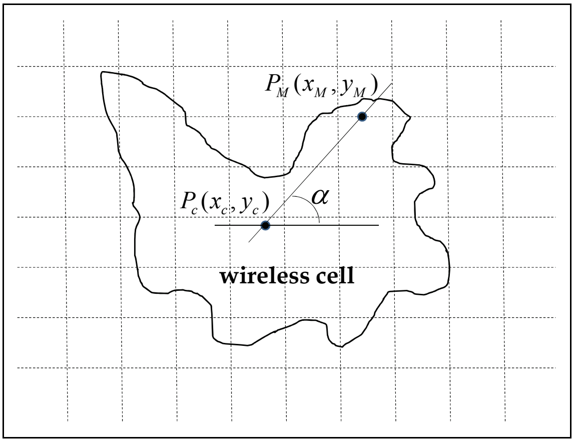


Fig. 10. Trigonometric approach for path loss map building in (anisotropic) wireless cell environment

The boundary of a cell in the power map is identified by a set of values of the power approaching the network sensitivity μ. When such values are not available in the power map they can be obtained through polynomial interpolation. The centre of the cell is instead identified with the maximum power value.

Once the boundary and centre of each cell *k*-th in the power map has been identified as a set of points, it is possible to assign an angle α_k to each point of the boundary with respect of the centre of the cell and its distance R_{cell} (α_k). The list of radius R_{cell} (α_k) for each cell *k*-th is exploited by the ELB scheme.

It follows that the maximum *GP* in a WLAN and UMTS cell can be calculated with the following approximated formulas, which replace (17) as:

$$\begin{cases} GP_{\max}^{UMTS} = BW_{\max}^{UMTS} \cdot \Pr\left\{\left(\dfrac{R_{cell}^{UMTS}(\alpha)}{r^{UMTS}}\right)^{\gamma} + \delta A_d < 1\right\} \\ GP_{\max}^{WLAN} = BW_{\max}^{WLAN} \cdot \Pr\left\{\left(\dfrac{R_{cell}^{WLAN}(\alpha)}{r^{WLAN}}\right)^{\gamma} + \delta A_d < 1\right\} \end{cases} \tag{29}$$

However, the handover decisions are still taken on the basis of (19).

4. Conclusions

In this chapter we described the main aspects of vertical handover procedure. This mechanism is oriented to ensure and maintain service continuity for mobile users in heterogeneous wireless network environments. Three different vertical handover strategies have been investigated, such as (*i*) the multi-parameter QoS-based approach, (*ii*) the location-based algorithm and (*iii*) the hybrid vertical handover technique.

The multi-parameter QoS-based VHO assumes both subjective and objective video quality metrics as handover decision criterion, such as a vertical handover is initiated whenever the QoS level is decreasing under a fixed threshold. In the location-based VHO the mobile node position is exploited in order to estimate some network performance (*i.e.* goodput figure). A handover is then initiated whenever a selected candidate network guarantees higher performance than the serving network.

Finally, we illustrated the third vertical handover technique (*i.e.* HVHO), that is an hybrid approach based on both power measurements and location information. The HVHO develops an enhanced location-based approach to build and maintain path loss maps, which provides an updated description of the wireless cells in a visited environment. The use of combined location and power information to drive handover decisions brings about goodput enhancements, while assuring a limited VHO frequency with respect to simple single-parameter techniques.

5. References

Makhecha K. P. & Wandra K. H. (2009). 4G Wireless Networks: Opportunities and Challenges, Annual IEEE India Conference (INDICON), pp.1-4, December 2009.

Lin M.; Heesook Choi; Dawson T. & La Porta T. (2010). Network Integration in 3G and 4G Wireless Networks, *Proceedings of 19th International Conference on Computer Communications and Networks* (ICCCN), pp.1-8, August 2010.

Balasubramaniam S. & Indulska J. (2004). Vertical handover supporting pervasive computing in future wireless networks, *Computer Communications*, Vol. 27, Issue 8, pp. 708–719, 2004.

Knightson K.; Morita N. & Towle T. (2005). NGN architecture: generic principles, functional architecture, and implementation, *IEEE Communication Magazine*, Vol. 43, Issue 10, pp. 49–56, October 2005.

McNair J. & Fang Z. (2004). Vertical handovers in fourth-generation multinetwork environments, *IEEE Wireless Communications*, Vol. 11, Issue 3, pp. 8–15, June, 2004.

Pollini G. P. (1996). Trends in handover design, *IEEE Communication Magazine*, Vol. 34, No. 3, March 1996, pp. 82–90.

Inzerilli T. & Vegni A. M. (2008). A reactive vertical handover approach for WiFi-UMTS dual-mode terminals, *Proceeding of 12th Annual IEEE International Symposium on Consumer Electronics*, April 2008, Vilamoura (Portugal).

Ayyappan, K. & Dananjayan, P. (2008). RSS Measurement for Vertical Handover in Heterogeneous Network, *Journal of Theoretical and Applied Information Technology*, Vol. 4, Issue 10, October 2008.

Vegni A. M.; Carli M.; Neri A. & Ragosa G. (2007). QoS-based Vertical Handover in heterogeneous networks, *Proceeding on 10th International Wireless Personal Multimedia Communications*, December 2007, Jaipur (India).

Yang K.; Gondal I.; Qiu B. & Dooley L. S. (2007). Combined SINR based vertical handover algorithm for next generation heterogeneous wireless networks, *Proceeding on IEEE GLOBECOM 2007*, November 2007, Washinton (USA).

Vegni A. M.; Tamea G.; Inzerilli T. & Cusani R. (2009). A Combined Vertical Handover Decision Metric for QoS Enhancement in Next Generation Networks, *Proceedings of IEEE International Conference on Wireless and Mobile Computing, Networking and Communications* 2009, pp. 233–238, October 2009, Marrakech (Morocco).

Kibria M. R.; Jamalipour A. & Mirchandani V. (2005). A location aware three-step vertical handover scheme for 4G/B3G networks, *Proceeding on IEEE GLOBECOM 2005*, Vol. 5, pp. 2752–2756, November 2005, St. Louis (USA).

Kim W. I.; Lee B. J.; Song J. S.; Shin Y. S. & Kim Y. J. (2007). Ping-Pong Avoidance Algorithm for Vertical Handover in Wireless Overlay Networks, *Proceeding of IEEE 66th Vehicular Technology Conference*, pp. 1509-1512, September 2007.

Inzerilli T.; Vegni A. M.; Neri A. & Cusani R. (2008). A Location-based Vertical Handover algorithm for limitation of the ping-pong effect, *Proceedings on 4th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, October 2008, Avignon (France).

Gupta V.; Williams M. G.; Johnston D. J.; McCann S.; Barber P. & Ohba Y. (2006) IEEE 802.21 Overview of Standard for Media Independent Handover Services, *IEEE 802 Plenary*, San Diego, CA, USA, July 2006.

Golmie N.; Olvera-Hernandez U.; Rouil R.; Salminen R. & Woon S. (2006). Implementing Quality of Service based handovers using the IEEE 802.21 framework, *IEEE 802.21 session 15* San Diego, California, July 2006.

Shin J.; Kim J. W. & Kuo C. C. J. (2001). Quality-of-Service Mapping Mechanism for Packet Video in Differentiated Services Network, *IEEE Transactions on Multimedia*, Vol. 3, no. 2, June 2001.

Vegni A. M. (2010). *Multimedia Mobile Communications in Heterogeneous Wireless Networks - Part 2*, PhD thesis, University of Roma Tre, March 2010, available online at http://www.comlab.uniroma3.it/vegni.htm

Wang S. S.; Green M. & Malkawi M. (2001). Adaptive Handoff Method Using Mobile Location Information, Proceedings on IEEE Emerging Technology Symp. Broadband Comm. for the Internet Era Symposium, pp. 97-101, September 2001.

Inzerilli T.; Vegni A. M.; Neri A. & Cusani R. (2010). A Cross-Layer Location-Based Approach for Mobile-Controlled Connectivity, *International Journal of Digital Multimedia Broadcasting*, vol. 2010, 13 pages, 2010.

12

On the Use of SCTP in Wireless Networks

Maria-Dolores Cano
Department of Information Technologies and Communications
Technical University of Cartagena
Spain

1. Introduction

Communications networks, particularly Internet, allow starting new businesses, to improve the current ones, and to offer an easiest access to new markets. Nowadays, Internet connects millions of terminals in the world, and it is a goal that this connection could be done with anyone, at any moment, and anywhere. In order to achieve this target, new lax and varied access requirements are needed. It is expected that a user would be able to access network services in a transparent way disregarding the location. The user terminal could seamlessly use the best available access technology (e.g., WLAN (Wireless Local Area Networks), LTE (Long Term Evolution), or PLC (Power Line Communications)), and service provisioning should agree with the user contract. This convergence of communications networks is giving rise to new challenges. The Internet Protocol (IP) has been selected to provide the necessary interconnection among all wireless and wired existing technologies. However, the use of IP does not solve all drawbacks. Multimedia applications show that current transport protocols like TCP (Transmission Control Protocol) or UDP (User Datagram Protocol) are not good enough to meet the new quality requirements.

To face these new challenges, the IETF (Internet Engineering Task Force) defined a new transport protocol called Stream Control Transmission Protocol (SCTP) (Stewart, 2007), whose main features are multihoming and multistreaming. Multistreaming allows transmission of several data streams within the same communication, splitting the application data into multiple streams that have the property of independently sequenced delivery, so that message losses in any one stream will only initially affect delivery within that stream, and not delivery in other streams. On the other hand, multihoming allows binding one transport layer's association to multiple addresses at each end of the SCTP association. The binding allows a sender to transmit data packets to a multihomed receiver through one of those different destination addresses. Therefore, SCTP is not only intended for signaling, but it can be used for any data application transport. The first studies about the performance of SCTP showed promising results. For instance, in (Kamal *et al.*, 2005), authors evaluate the benefits of using SCTP instead of TCP as the underlying transport protocol for a MPI (Message Passing Interface) middleware. Darche *et al.* (2006) presented a network architecture to enhance the cooperation of mobile and broadcast networks using SCTP as the transport layer protocol. In (Shaojian *et al.*, 2005), authors study the suitability of SCTP for satellite networks. Kim *et al.* investigate in (Kim *et al.*, 2006) the applicability of SCTP in MANET (Mobile Ad hoc NETworks). In (Kozlovszky *et al.*, 2006), authors carry out performance measurements with TCP and SCTP as protocols to be used in distributed cluster environments. Finally, in (Natarajan *et al.*, 2006) authors propose the use of SCTP for HTTP-based applications, showing the benefits with real web servers compatible with SCTP. All these works showed the notable performance of SCTP as a multipurpose transport layer protocol.

This chapter reviews the specific use of SCTP in wireless networks and illustrates how to implement a multipurpose SCTP client/server application, compatible with IPv6, from a practical point of view. We describe how to enable multistreaming and multihoming capabilities. Through experimental tests in wired and wireless networks, we measure the SCTP performance regarding multistreaming and multihoming operation, compare it with the TCP protocol, and discuss its advantages and drawbacks. Therefore, the main contribution of this chapter is to present a survey in the work carried so far to turn the SCTP into a feasible transport-protocol option for wireless networks and to show the practical aspects of the design of a SCTP's open source client/server application, including some basic, but explanatory, experimental results in a single server – single client scenario. This work reveals that SCTP may be a competitive transport protocol for multimedia applications.

The rest of the chapter is organized as follows. Section 2 reviews the SCTP characteristics and its applicability in wireless networks. Section 3 explains how to make a SCTP client/server application. Experimental results are shown and discussed in Section 4. The chapter ends with conclusions in Section 5.

2. Related work

The SCTP features are described in this section. In addition, a survey about the applicability of SCTP in wireless environments has been also included. Among the advantages of using SCTP in wireless networks, mobility and multimedia transmission are highlighted, reviewing the most relevant works in these two areas. Other improvements like security or the introduction of redundancy for data delivery are also mentioned.

2.1 Stream control transmission protocol

SCTP is a message oriented transport protocol. Like TCP, SCTP provides a reliable transport service ensuring that data arrives in sequence and without errors. Like TCP, SCTP is a session-oriented mechanism, meaning that a relationship is created between the endpoints of a SCTP association prior to data being transmitted, and this relationship is maintained until all data transmission has been successfully completed. However, SCTP includes some new features (see Table 1) that evidence the advantages of using it in applications needing transport with additional performance and reliability.

Multihoming. A SCTP endpoint has the ability to work with more than one IP address, thus a session can remain active even in the presence of network failures. One of the main advantages is that in a conventional single-homed session, the failure of a local area network access can isolate the end system, but with multi-homing, redundant local area networks can be used to reinforce the local access. Multi-homing is not used for redundancy, as indicated in (Stewart, 2007). A pair of IP addresses <source, destination> is defined as the primary path, being used for data transmission. The other combinations of source and destination addresses will be considered as alternative paths, and will be employed in case of a primary path failure, which is detected by using the heartbeat mechanism (monitoring function). The IP addresses of the SCTP association could be exchanged even if the association is already in use, i.e., it is possible to include new IP addresses during the communication (Stewart *et al.*, 2007). This feature is known as Dynamic Address Reconfiguration or Mobile SCTP.

Characteristics	**TCP**	**UDP**	**SCTP**
Unicast	Yes	Yes	Yes
Byte oriented	Yes	No	No
Message oriented	No	No	Yes
Reliable transport service	Yes	No	Yes
Multi-homing	No	No	Yes
Multi-stream	No	No	Yes
Cookie mechanisms	No	No	Yes
Rate adaptive	Yes	No	Yes
Heartbeat mechanism	No	No	Yes

Table 1. TCP, UDP, and SCTP comparison

Heartbeat Mechanism. A SCTP source should check if it is possible to reach the remote endpoint. This is done by means of the heartbeat mechanism. Alternative paths are monitored with heartbeat messages. Heartbeat messages are small messages with no user-data periodically sent to the destination addresses, and immediately acknowledged by the destination. The sender of a heartbeat message should increment a respective error counter of the destination address each time a heartbeat is sent to that address and not acknowledged within the corresponding time interval (RTO, Retransmission TimeOut). If this counter reaches a maximum value, the endpoint should mark this address as inactive. On the contrary, upon the receipt of a heartbeat acknowledgement, the sender of the heartbeat should clear the error counter of the destination address to which the heartbeat was sent, and mark the destination address as active.

Multistreaming. This feature allows splitting the application data into multiple streams that have the property of independent sequenced delivery, so that message losses in any one stream will only initially affect delivery within that stream, and not delivery in other streams. This is achieved by making independent data transmission and data delivery. SCTP uses a Transmission Sequence Number (TSN) for data transmission and detection of message losses, and also a Stream ID/Stream Sequence Number pair, which is used to determine the sequence of delivery of received data. Therefore in reception, the end point can continue to deliver messages to the unaffected streams while buffering messages in the affected stream until retransmission occurs.

Initiation. SCTP initiation procedure requires four messages. A cookie mechanism was incorporated to avoid Denial of Service (DoS) attacks. A SCTP client sends an init message to the SCTP server. The server replies with an init ack message that includes a cookie (a TCB (Transmission Control Block), a validity period, and a signature for authentication). Since the init ack is addressed to the source IP address of the init message, an attacker cannot get the cookie. A valid SCTP client would get the cookie, and send it back in a cookie echo message to the server. When this packet is received, the server starts giving resources to the client. The procedure finishes with a cookie ack message.

Data Exchange. Data exchange in SCTP is very similar to the TCP SACK procedure (Stewart, 2007). SCTP uses the same congestion and stream control algorithms as TCP.

Shutdown. SCTP shutdown procedure uses three messages: shutdown, shutdown ack, and shutdown complete. Each endpoint has an ack of the data packets received by the remote endpoint before closing the connection. SCTP does not support a half open connection, but it is assumed that if the shutdown initiates, then both endpoints will stop transmitting data.

2.2 SCTP in wireless networks

Seamless mobility is one of the challenges in wireless networks. With the proliferation of new types of wireless access technologies (e.g., WiFi, WiMAX, 3G, vehicular networks, etc.), a user, through his/her mobile device, should be able to change his/her location maintaining the Quality of Service (QoS) performance disregarding the roaming, either horizontal (under the same technology) or vertical (crossing different technologies). SCTP is a competitive solution for mobility due to its multihoming capability. Multimedia transmission is another challenge in wireless networks due to the higher likelihood of packet losses (error-prone channels). In this case, SCTP multistreaming improves the data rate throughput since streams are independently delivered; hence, the multimedia application is less sensitive to packet losses. Finally, some new modifications to SCTP have been presented in the related literature to increase its performance, e.g., allowing redundancy in multihomed devices. This section reviews the most relevant works in these areas.

2.2.1 Mobility and handovers in wireless networks

Several works in the related literature had demonstrated the advantages of using SCTP to improve both vertical or horizontal handovers and signaling in wireless networks. Authors in (Afif *et al.*, 2006a) proposed to include a new type of chunk in SCTP able to send QoS transmission parameters over the radio interface from an EGPRS mobile to the SCTP peer. By doing so, SCTP could adapt the transmission rate depending on the radio transmission conditions (e.g., LLC error rate, RLC/MAC block error rate, etc.). The reason to incorporate this new chunk, as stated by the authors, can be explained as follows. Even though SCTP is able to change the IP addresses in use, data packets are sent to old IP address before the alternative addresses become the primary ones. Therefore, there are packet losses during the exchange process. The simulation study in an EGPRS network with handovers between cells showed that the achieved throughput is higher with this modification than with the standard SCTP implementation because fewer packets are lost during handovers. From a similar perspective, same authors verified in (Afif *et al.*, 2006b) that their modification is also useful for handovers between EGPRS and Wireless Local Area Networks (WLAN).

Honda *et al.* proposed a new handover mechanism based on SCTP and a new data retransmission feature for smooth handover. In their work, authors state that the exchange of addresses in SCTP, assuming the new addresses to use are unknown at the beginning of the SCTP association (i.e., using Dynamic Address Reconfiguration), suffers a high delay mainly due to the multiple RTO expirations required to identify the failure. To overcome this situation, authors propose to include two algorithms called FastAssociation Reconfiguration and Fast Transmission Recovery. The former minimizes the RTO needed to substitute the addresses in use, whereas the latter allows sending data just after the establishment of the new addresses. Observe that in the standard, it was necessary to wait an RTO after a new path is configured to send data. The evaluation, carried out in an experimental network with WLAN links, showed that the handover latency was notably reduced using the authors' approach.

Focusing on vertical handover between WLAN and cellular networks, particularly UMTS (Universal Mobile Telecommunication System), authors in (Ma *et al.*, 2007) proposed a very interesting error recovery scheme called Sending-buffer Multicast-Aided Retransmission with Fast Retransmission that increases the throughput achieved during the SCTP connection in the presence of forced vertical handovers from WLAN to UMTS. A forced vertical handover occurs when the mobile node leaves the WLAN coverage due to the loss of signal and switches to the cell network. The advantages of using SCTP for vertical handovers were clearly identified in (Ma *et al.*, 2004): higher throughput, shorter delay, a simpler network architecture, and ease to adapt network congestion and flow control parameters to the new network; but a scenario with forced handovers involves important packet losses. Ma, Yu & Leung (2007) categorized these packet losses as dropping consecutive packets because of the loss of signal (WLAN) and random packet losses over the cellular link. To deal with these different types of errors, the authors propose to use two solutions. First, packet losses due to the loss of signal enable the Sending-buffer Multicast-Aided Retransmission algorithm, which multicast all buffered data on both the primary and the alternate address (observe that in a standard implementation SCTP only retransmits data to the alternate address if the error was due to a time out). The same applies to new data that needs to be sent. Second, packet losses likely due to random packet losses over the link (detected by the reception of duplicated acknowledgments) activate the Fast Retransmission algorithm, which force the retransmission to be done to the same destination IP address. With these two algorithms, long waiting delays are avoided, thus increasing the achieved throughput. Working on the same heterogeneous scenario with WLAN and UMTS networks, Shieh *et al.* (2008) detected that SCTP significantly decreases the congestion window when new primary addresses are used in the SCTP association (i.e., during a handover). Therefore, they proposed to assign an adequate initial congestion window according to the bandwidth available in the new path, so the association can skip the slow-start phase and enter the congestion avoidance phase directly. Packet-pair bandwidth proving is used to estimate the available bandwidth in the new path. Authors demonstrated the feasibility and goodness of their proposal through simulation. From an experimental point of view, authors in (Bokor *et al.*, 2009) designed and implemented a real native IPv6 UMTS-WLAN testbed to evaluate the effect of SCTP parameter configuration in terms of handover effectiveness, link changeover characteristics, throughput, and transmission delay. Among the most important parameters that have an effect on handover are: *RTO.Min*, *RTO.Max*, *Path.Max.Retransmission*, and *HB.Interval*. Authors verified that with the standard parameters, the handover delay would rise exponentially due to RTO redoubling, but using a more appropriate setting the handover delay rises linearly when the RTO is incremented. They also recommended keeping the *HB.Interval* (the time that elapses between consecutive heartbeat monitoring messages) as low as possible. Finally, they found that the SCTP performance in terms of delay, jitter, and throughput was better in UMTS than in WLAN.

From another perspective, authors in (Lee *et al.*, 2009) studied a mobile web agent framework based on SCTP. Typical web agents use TCP as transport protocol. However, mobile web agents using TCP present the following drawbacks: performance degradation, head-of-line (HOL) blocking, and unsupported mobility (as identified by IEEE Std 802.11-1997 and IEEE 802.16e-2005). By transmitting each object in a separate stream, SCTP solves the HOL problem. Mobility is achieved by the SCTP multihoming capability. To improve the performance, authors assumed that mean response time between HTTP requests and replies is the most important performance parameter in a web environment. Therefore, they proposed to use SCTP to decrease the response time compared to the classical TCP implementation of web agents. Authors described the complete architecture for the mobile SCTP web agent framework. By simulation, they found that the mean response time decreased notably (around 30%) by using SCTP. The mean packet loss was also smaller with SCTP, and the faster the moving speed the better the SCTP performance in terms of packet loss compared to TCP.

Regarding the option of introducing crosslayer techniques to combine the SCTP features with information available at lower levels, the IEEE introduced the IEEE 802.21-2008 Media Independent Handover (MIH) as a way to provide link layer intelligence and other related network information to upper layers. MIH does not carry out the network handover, but it provides information to allow handover within a wide range of networks (e.g., WiFi, WiMAX, 3G, etc.). In (Fallon *et al.*, 2009) authors proposed to separate path performance evaluation (i.e., how SCTP detects that a path is no longer available) from path switching (i.e., update the new addresses of the primary path in the SCTP association). Whereas the first task will be done with MIH, SCTP will only be in charge of the second task (path performance is disabled in SCTP). By simulation, authors demonstrated that the combination of SCTP and MIH reacts to sudden performance degradation resulting from obscured line of sight in a heterogeneous scenario with WiMAX and HSDPA technologies. Indeed, the throughput of the SCTP connection improved notably (from 5% to 45%) compared to the standard SCTP implementations.

Network Mobility (NEMO), commonly used in military or vehicular applications, has been also studied from a SCTP perspective. In host mobility, a network in which terminals change their location, mobility is managed through the mobile node itself. In a mobile network, mobility is managed by a central node (e.g., a bus providing a WLAN service that moves around a city, hence changing the access point from which obtains Internet access). Leu & Ko (2008) proposed a method that combines SIP and SCTP with the aim of minimizing delay and packet losses during the handovers of a mobile network. With the authors' proposal, packet losses decreased significantly. Similarly, Huang & Lin (2010) presented a method to improve the bandwidth use and the achieved throughput in vehicular networks by using SCTP. Their approach is explained as follows. In a Vehicle to Infrastructure network (V2I), moving vehicular nodes communicate with Road Side Units (RSU) deployed in a specific area. RSU are connected to the wired infrastructure, e.g., providing Internet access to mobile vehicular nodes. Usually, several RSU share the same gateway to access the infrastructure. Therefore, authors proposed to use this gateway as a SCTP-packet monitoring station, buffering all SCTP packets containing data chunks. In the event of a packet loss, the gateway (not the destination node, which is assumed to be in the wired part of the network) will be in charge of retransmitting lost packets in the wireless link. With this scheme, the wired part of the communication is used more efficiently because no retransmissions are sent (unless the packet loss occurs in the wired part of the network). Moreover, since the destination node is not informed about packet losses in the wireless part of the network, its congestion window does not decrease as much, keeping a higher throughout rate in average. The performance of this proposal was done through simulation. Authors verified that the achieved throughput, the transmission time, and the congestion window behaved better with their approach than with the standard SCTP implementation.

2.2.2 Multimedia transmission over wireless networks

The use of multimedia services and applications over wireless links is another important research area. Authors in (Wang *et al.*, 2003) presented one of the first works evaluating the performance of Partial Reliability SCTP (PR SCTP), a modification of SCTP that provides unreliable transmission service to part of the data to be sent, as the transport protocol for video (MPEG-4) transmission in a wireless local area network. Results showed an improvement in the video quality comparing PR SCTP with UDP. Another interesting works regarding MPEG-4 video transmission over wireless technologies are presented in (Nosheen *et al.*, 2007) and (Chughtai *et al.*, 2009). In the first work, authors compared SCTP with UDP and DCCP (Datagram Congestion Control Protocol) (Kohler *et al.*, 2006). By simulation, they found that the throughput achieved by UDP could be more than 20% smaller than the throughput achieved by SCTP or DCCP in a wireless environment. However, the delay was higher in SCTP due to the congestion control mechanism. In the presence of background traffic, the results also showed that SCTP and DCCP outperformed UDP. As an extension to this work, Chughtai *et al.* (2009) carried out a similar study to compare the QoS performance of SCTP, UDP, and SCTP transmitting video in a WiMAX network. The simulation scenarios included downloading or uploading MPEG-4 video traffic using a different number of subscribers, different packet sizes, and a variable video rate. Results showed that delay and jitter were lower with SCTP than with UDP or DCCP. In terms of throughput, DCCP performed slightly better than SCTP, and both exceeded UDP performance.

Wang *et al.* (2008) also studied video delivery over wireless networks using SCTP. They focused on the multistreaming feature of SCTP, and how to use it to optimize video quality. Previous works from the literature such as (Balk *et al.*, 2002) showed the benefits of using multistreaming for MPEG-4 video transmission in wired network by applying a differential treatment among streams in a SCTP association. Differing from previous works, Wang *et al.* (2008) proposed MPEG-4 transmission with optimized partial reliability among streams in a heterogeneous scenario with error-prone 802.11 wireless channels. Their proposal was based on retransmitting packets belonging to stream of I-frames until packets are eventually received, while no retransmissions are attempted for packets in stream of B- and P- frames. In terms of retransmission overhead delay, simulation results showed that adjusting SCTP fast retransmit threshold can reduce the retransmission overhead delay, hence increasing the I-frame data rate, and the video quality. Furthering the results obtained in this work, the same authors introduced in (Wang *et al.*, 2009) an extension to the SCTP protocol. The goal was to improve the transmission of delay sensitive multimedia data by including a selective retransmission of lost packets depending on whether the lost packets would still arrive before the schedule time. Assuming that there is clock synchronization between the SCTP associated peers, authors included a new field to the SCTP header with the time a packet is sent, so that the endpoint after reception can estimate the one-way delay. This value is sent to the sender from the receiver in the acknowledgement packet. Then, in the receiver side, the time of each frame of MPEG-4 to be played out is calculated, so if the frame is not received before this schedule time will be considered as non-useful and its retransmission will not be necessary. By simulation, authors achieved interesting results, confirming the improvement in the MPEG-4 video transmission performance.

Voice over IP (VoIP) is another important application that is gaining momentum. Chang *et al.* (2009) presented a middleware to transfer the session initiation protocol (SIP) signaling and real-time transmission protocol (RTP) messages from using UDP or TCP to SCTP.

Switching from UDP or TCP to SCTP (with Dynamic Address Reconfiguration) provides a seamless way for the user to roam maintaining the QoS level of the VoIP call. Authors analyzed their proposal in a real testbed. Nevertheless, results showed that although mobility was achieved, the delay was higher with their proposal.

Live TV broadcasting over wireless technologies could also benefit from the use of SCTP. Liu *et al.* (2010) introduced a method to provide an economic way of live news broadcasting by using SCTP. Satellite News Gathering (SNG) vehicles, which usually use satellite links for transmission, are an expensive service for TV companies, mainly due to the required equipment. In this case, the current deployment of WiMAX networks is a feasible alternative to satellite communication, but the bandwidth offered by WiMAX is not enough to provide a live TV service with QoS demands. Therefore, the authors proposed to take advantage of all available wireless networks, not only WiMAX but also HSDPA or WiFi, thus increasing the available bandwidth. A SCTP multi-link connection with both multihoming and multi-streaming was a key point for this implementation. SCTP Concurrent Multipath Transfer, which will be explained in next section, is also needed. With an experimental testbed, authors demonstrated the feasibility of their proposal, not only achieving a cost-effective system to provide live TV broadcasting but also increasing the coverage of previous SNG systems.

2.2.3 Other SCTP improvements

Concurrent Multipath Transfer (CMT) consists of simultaneously sending data over all available paths, hence, increasing the bandwidth of the SCTP association (Iyengar *et al.*, 2006). In environments where the paths of the SCTP association exhibit very different network conditions (e.g., round trip times or bandwidth), packet reordering is required in the receiver side, and this might cause retransmission, lowering the connection rate. To avoid this situation, authors in (Perotto *et al.*, 2007) compared the performance of two SCTP modifications: Sender-Based Packet Pair SCTP (SBPP-SCTP) and Westwood SCTP (W-SCTP). The former uses the sender-based packet pair technique, mentioned in the previous section, to estimate the bottleneck bandwidth of each path. The latter uses the same algorithm as in TCP Westwood (Mascolo *et al.*, 2004) for the bandwidth estimation. Both aim at minimizing packet reordering. In presence of intermittent interfering cross-traffic, authors showed that W-SCTP achieves a higher throughput than SBPP-SCTP. Aydin & Shen (2009) studied the performance of CMT SCTP over 802.11 static multihop wireless networks. They compared CMT SCTP with three different techniques: i) standard SCTP using just one path (the best one in terms of bandwidth) to send data, ii) standard SCTP using just one path (the worst one in terms of bandwidth) to send data, and iii) standard SCTP using all available paths to send data (splitting the traffic into the different available paths of the SCTP association). Results showed that in a multihop wireless scenario the achieved throughput is higher with CMT SCTP than with any of the three alternatives used for comparison. Nevertheless, CMT SCTP still presents a drawback to be completely useful for wireless networks: the received buffer blocking problem. This problem was clearly stated in (Wang *et al.*, 2010): "In SCTP transmission, data streams between each other are logically independent, if receiver has received all data chunks of a certain stream. The data of this stream can be delivered to the application layer. But in traditional CMT, because data chunks of the same stream maybe transferred to different paths, the data chunks could not arrive at the receiver orderly and duly, so the receive buffer blocking problem happens. This problem can seriously influence network performance, especially in high error rate and delay wireless network." Consequently, authors proposed a new modification of the SCTP called Wireless Concurrent Multipath Transfer SCTP (WCMT SCTP). With this modification, each SCTP path delivers packets belonging to the same stream (one or more than one). For instance, if there are three paths available and there are five streams, then the first path only transmits packets from the first stream, the second path only transmits packets from streams two and three, and the third path only transmits packets from streams four and five. Authors also added other changes to the standard CMT implementation: a per-path congestion control mechanism, a new congestion control mechanism and a new retransmission mechanism that takes into account the type of error. Results obtained by simulation showed that WCMT SCTP performs better than CMT SCTP in ad hoc networks. In a similar way, Yuan *et al.* (2010) improved the CMT SCTP mechanism by categorizing the streams depending on their specific QoS requirements, and grouping those streams with similar QoS needs in subflows that are sent through the more appropriate paths available in the SCTP association. Finally, the work done in (Xu *et al.*, 2011) showed how to optimize CMT SCTP for video and multimedia content distribution.

Another interesting works that improve the performance of SCTP in wireless environments from different perspectives are (Cui *et al.*, 2007; Cano *et al.*, 2008; Lee & Atiquzzaman, 2009; Cheng *et al.*, 2010; Funasaka *et al.*, 2010). Cui *et al.* (2007) proposed to use a hierarchical checksum method that improves the retransmission procedure, thus increasing the achieved throughput in links with high packet losses. Cano *et al.* (2008) investigated how to combine the use of IPSec (Internet Protocol Security) with SCTP to enhance the security of the wireless communication. The work done in (Lee & Atiquzzaman, 2009) presented an analytical model to estimate the delay of HTTP over SCTP in wireless scenarios. Last, Cheng *et al.* (2010) proposed to use two new methods for bandwidth estimation and per-stream-based error recovery.

Library	Description
netinet/sctp.h	It contains definitions for SCTP primitives and data structures.
netdb.h	It contains definitions for network database operation, e.g. translation.
sys/socket.h	It defines macros for the Internet Protocol family such as the datagram socket or the byte-stream socket among others.
netinet/in.h	It contains definitions of different types for the Internet Protocol family, e.g. sockaddr_in to store the socket parameters (IP address, etc.).
arpa/inet.h	To manage numeric IP addresses, making available some of the types defined in netinet/in.h

Table 2. Description of the libraries related to SCTP network communication

3. Implementation

For the sake of simplicity, we implement three SCTP client/server applications. The first one is called single SCTP, the second one is called multistream SCTP, and the last one is called multihomed SCTP. Single SCTP is very similar to TCP, since it will be able to transmit just one data stream between source and destination endpoints. Multistream SCTP includes multistreaming, and finally, multihomed SCTP incorporates multihoming. The three implementations are written in C code. We use the libraries provided by the Berkeley Socket Application Programming Interface, which are briefly described in Table 2. Next sections detail the practical SCTP implementation issues.

```
1  int main(int argc, char *argv[])
2  {
3   int sockfd;
4   struct hostent *host;
//Structures to manage IP address
5   struct sockaddr_in remote_addr;   //IPv4
6   host = gethostbyname(argv[1]);
7   ra_family = host->h_addrtype;  //AF_INET
//IPv4 socket
8   sockfd = socket( ra_family, SOCK_STREAM, IPPROTO_SCTP);
9  if(sockfd == -1)
10    {perror("Socket:");exit(1);}
//Set server IP address
11  remote_addr.sin_family=AF_INET;
12  remote_addr.sin_port=htons(REM_PORT);
13  remote_addr.sin_addr=*((struct in_addr *)host->h_addr);
14  bzero(&(remote_addr.sin_zero),8);
//Connect to server
15  if(connect(sockfd,(struct sockaddr*)&remote_addr,sizeof(struct sockaddr))==-1)
16    {perror("connect:"); exit(1);}
//Omitting lines of code to receive a file
//Close socket
17  close(sockfd);
18  return 0;
19 }
```

Fig. 1. Extract of the original SCTP client code in a single file transmission

3.1 Single SCTP

SCTP server and SCTP client structures are very similar to those used in TCP. Fig. 1 shows how to implement a SCTP client. The only difference with TCP is in the *socket()* function, where the protocol type field should be IPPROTO_SCTP instead of the common parameter 0 used for TCP or UDP transport protocols (see code line 8 in Fig. 1). The rest of the implementation is done as in TCP; i.e., once the socket is created, the server IP address is set (see code lines 11-14 in Fig. 1), and the client connects to the server (see code line 15 in Fig. 1). Observe that we use the server IP address as an argument in the command line (see code line 6 in Fig. 1). If we want to use IPv6 instead of IPv4, some simple changes included in Table 3 are needed. First, it is necessary an appropriate structure to store an IPv6 address. Second, the *gethostbyname()* function, needs to know that the IP address is an IPv6 one, and the same applies to all lines of code where we use the IP address.

In Fig. 2, we define how to implement a SCTP server. In this case, to execute the server, no parameters are needed in the command line. We define a constant called MYPORT to include the port number associated to the server IP address (see code line 12 in Fig. 2). The server IP address is automatically set to any local IP address available (see code line 13 in Fig. 2). Then, we follow the usual sequence to set up the server. First, we create the socket with the *socket()* function. As indicated before, the socket protocol is set to IPPROTO_SCTP (see code line 15 in Fig. 2). Then, we set the socket parameters with the *bind()* function (see code line 18 in Fig. 2). Afterwards, we execute *listen()* so that the server can receive a specific number of client requests (see code line 20 in Fig. 2). The *accept()* function makes the server to wait for client requests (see code line 25 in Fig. 2). Finally, if a client request is received, the client is served by a child process due to the *fork()* function (see code line 27 in Fig. 2). To make it compatible with IPv6, lines indicated in Table 4 should be replaced.

```
1  int main(int argc, char *argv[])
2  {
3   int sockfd, newfd;
4   socklen_t sin_size;
5   struct sockaddr_in local;
6   struct sockaddr_in remota;
7   struct hostent *host;
8   sa_family_t la_family;
9   la_family = host->h_addrtype;
10  host = gethostbyname(argv[1]);
11  local.sin_family = AF_INET;
12  local.sin_port = htons(MY_PORT);
13  local.sin_addr.s_addr = htonl(INADDR_ANY);//Any local IP address
14  bzero(&(local.sin_zero),8);
15  sockfd = socket( la_family, SOCK_STREAM, IPPROTO_SCTP);
16  if(sockfd == -1)
17   {perror("Socket:"); exit(1);}
18  if((bind(sockfd, (struct sockaddr*)&local, sizeof(struct sockaddr)))==-1)
19   {perror("bind");exit(1);}
20  if(listen(sockfd,5) == -1)
21   {perror("listen");exit(1);}
22  for(;;)
23  {
24   sin_size=sizeof(struct sockaddr_in);
25   if((newfd = accept(sockfd, (struct sockaddr*)&local,&sin_size)) == -1)
26     {perror("accept");exit(1);}
27   if (!fork()
//Omitting lines of code to send a file
28  while(waitpid(-1,NULL,WNOHANG)>0);}
```

Fig. 2. Extract of the original SCTP server code in a single file transmission

Line# in Fig. 1	New code for IPv6

5	struct sockaddr_in6 remote_addr6; //IPv6
	struct in6_addr ipv6; //To store IPv6 address

6	host = gethostbyname2(argv[1], AF_INET6); //get IP address

8 to 14	sockfd = socket(ra_family, SOCK_STREAM, IPPROTO_SCTP);
	if(sockfd == -1) {perror("Socket:");exit(1);}
	remote_addr6.sin6_family = AF_INET6;
	remote_addr6.sin6_flowinfo = 0;
	remote_addr6.sin6_port = htons(REM_PORT);
	inet_pton(AF_INET6, argv[2], ipv6.s6_addr);
	remote_addr6.sin6_addr = ipv6;

15	if(connect(sockfd,(struct sockaddr*)&remote_addr6,sizeof(struct sockaddr))==-1)

Table 3. How to make the SCTP client implementation compatible with IPv6. Lines indicated in the first column should be replaced with lines shown in the second column

Line# in Fig. 1	New code for IPv6

5-6	struct sockaddr_in6 local6;
	struct sockaddr_in6 remota6;

10	host = gethostbyname2(argv[1], AF_INET6);

11 to 14	local6.sin6_family = AF_INET6;
	local6.sin6_flowinfo = 0;
	local6.sin6_port = htons(MY_PORT);
	local6.sin6_addr = in6addr_any;

18	if((bind(sockfd, (struct sockaddr*)&local6, sizeof(struct sockaddr))) == -1)

24-25	sin_size=sizeof(struct sockaddr_in6);
	if((newfd = accept(sockfd, (struct sockaddr*)&local6,&sin_size)) == -1)

Table 4. How to make the SCTP server implementation compatible with IPv6

3.2 Multistream SCTP

A SCTP client/server application with multistreaming allows sending/receiving multiple streams simultaneously. For instance, these different streams could belong to different files, so it would be possible transferring several files with the same SCTP association. Thus, the client only uses one request to the SCTP server. Nowadays, file downloading (music, games, software, etc.) is one of the most important services driving the usage of Internet. With the multistreaming SCTP feature, a unique association between SCTP client and SCTP server may accept many multimedia file transmissions, resulting in bandwidth saving as it will be shown in Section 4. The less traffic in the network, the more efficient the use.

From the SCTP client point of view, the multistream operation has to be enabled by setting some particular properties. The sequence is as follows. First, we create all data structures. Second, we create the SCTP socket as explained in the previous section (see code line 10 in Fig. 3). Then, the maximum number of ingoing and outgoing streams should be indicated. Accordingly, *setsockopt()* is used to set the number of flows or streams in the client/server

```
1   int main()
2   {
3   int connSock, in, ret;
4   struct sockaddr_in servaddr;
5   struct sctp_status status;
6   struct sctp_sndrcvinfo sndrcvinfo;
7   struct sctp_event_subscribe events;
8   struct sctp_initmsg initmsg;
9   int numElem=0, firstTime=1;
10  connSock = socket( AF_INET, SOCK_STREAM, IPPROTO_SCTP ); //IPv4
11  memset( &initmsg, 0, sizeof(initmsg) );
12  initmsg.sinit_num_ostreams = 30; //max streams
13  initmsg.sinit_max_instreams = 30; //max streams
14  initmsg.sinit_max_attempts = 5; //max attempts
15  ret = setsockopt( connSock, IPPROTO_SCTP, SCTP_INITMSG,&initmsg,
    sizeof(initmsg) );
16  bzero( (void *)&servaddr, sizeof(servaddr) ); //server to connect to
17  servaddr.sin_family = AF_INET;
18  servaddr.sin_port = htons(MY_PORT_NUM);
19  servaddr.sin_addr.s_addr = inet_addr("192.168.1.10" );
20  ret = connect( connSock, (struct sockaddr *)&servaddr, sizeof(servaddr) ); //connect
    to the server
21  memset( (void *)&events, 0, sizeof(events) );
22  events.sctp_data_io_event = 1;
23  ret = setsockopt( connSock, SOL_SCTP, SCTP_EVENTS,(const void *)&events,
    sizeof(events) );
24  //File transfer
25  //Loop to receive the different streams
26  in = sctp_recvmsg( connSock, (void *)buffer, sizeof(buffer),(struct sockaddr *)NULL,
    0, &sndrcvinfo, &flags );
27  if(in==0)
28     break;
29  //Store each stream in its corresponding file
30  if (sndrcvinfo.sinfo_stream == STREAM1)
31     { if(firstTime)
32         {fp=fopen("reciboweb.txt","wb"); firstTime=0;}
33     numElem=fwrite(buffer, 1, 1024, fp);
34     if(num_elementos<1024) {fclose(fp); break;} }
35  else if (sndrcvinfo.sinfo_stream == STREAM2) 42
36  {//Save this stream in its corresponding file, lines 31-41}
37  else if ...//Save each stream in its corresponding place
38  //End loop to receive different streams
39  fclose(fp);
40  close(connSock);
41  return 0; }
```

Fig. 3. Extract of the original SCTP client code in a multistream transmission

SCTP association. Both client and server agree on this parameter (see code lines 12-15 in Fig. 3). Afterwards, both the server IP address and the port number to connect to are indicated (see code lines 16-19 in Fig. 3). Next, the client connects to the server (see code line 20 in Fig. 3). Finally, we enable data delivery with the function *setsockopt()* (see code lines 21-23 in Fig. 3). By doing so, the client is able to use the primitive *sctp_recvmsg()* for data delivery. At this point, the client is ready to receive data in multiple streams within the same SCTP association.

On the other hand, the SCTP server multistream implementation also needs some variations compared to the SCTP server single implementation. First, we declare data structures. After that, we create the SCTP socket as explained in the previous section (see code line 8 in Fig. 4). Then, the server IP address is automatically set to any local IP address available, the port is assigned, and the *bind()* function is called (see code line 9-13 in Fig. 4). The maximum number of ingoing and outgoing streams is specified now (see code lines 15-18 in Fig. 4). Observe that it is the same value used previously for the client implementation. Next, the server remains listening for client requests (see code line 19 in Fig. 4). If there is a client request, then the server accepts the connection, and it starts sending the corresponding files.

Once a client is connected to the server, the information sent from the server to the client should be identified, so that the client knows what file (stream) the data belong to. Whereas the source (the server in this case) is in charge of assigning an identifier to each stream, which is done with the *sctp_sendmsg()*function and a stream number (see line 28 and 35 in Fig. 4), each stream is identified using the *sndrcvinfo.sinfo_stream* field (see line 30 and 35 in Fig. 3) in the receiving side (the client in this case).

3.3 Multihomed SCTP

In this section, we describe the additional code necessary to facilitate the SCTP multihomed feature. After calling the *bind()* function and before the SCTP association is established, any additional address should be enabled. Otherwise, multihoming cannot be used unless Dynamic Address Reconfiguration is set. Enabling addresses is done with the *sctp_bindx()* function. *sctp_bindx()* links any IP address (IPv4 or IPv6) to the SCTP association. It can be also used to delete an IP address from an association. Table 5 shows the new lines of code.

New code for multihoming

hst_adicional = gethostbyname(argv[3]);//get additional address/es
sctp_bindx(sockfd, (struct sockaddr*)ip4, 1, SCTP_BINDX_ADD_ADDR);

Table 5. How to make the SCTP client/server implementation with multihoming

4. Experimental results

Three different scenarios are evaluated to compare the performance of SCTP vs. TCP. In the first scenario, our SCTP application transfers a single text file, a single mp3 file, or a single mpeg file from server to client. We called it the single operation. In the second scenario, our SCTP application transmits different types of files simultaneously from server to client. We called it the multistream operation. The former is like a normal TCP transfer file operation. The latter could emulate a web loading, where usually different types of multimedia files are involved. In the third scenario, we test the multihoming feature in what we called the multihomed operation. Next we describe the experimental topology, and discuss the experimental results.

```
int main()
{
int listenSock, connSock, ret, msglen;
struct sockaddr_in servaddr;
struct sctp_initmsg initmsg;
FILE *fp;
int num_bytes=0;
listenSock = socket( AF_INET, SOCK_STREAM, IPPROTO_SCTP );
bzero( (void *)&servaddr, sizeof(servaddr) );
servaddr.sin_family = AF_INET;
servaddr.sin_addr.s_addr = htonl( INADDR_ANY );
servaddr.sin_port = htons(MY_PORT_NUM);
ret = bind( listenSock, (struct sockaddr *)&servaddr, sizeof(servaddr) );
memset( &initmsg, 0, sizeof(initmsg) );
initmsg.sinit_num_ostreams = 30;
initmsg.sinit_max_instreams = 30;
initmsg.sinit_max_attempts = 5;
ret = setsockopt( listenSock, IPPROTO_SCTP, SCTP_INITMSG, &initmsg, sizeof(initmsg) );
listen( listenSock, 5 );
int i=0;
while( 1 )
{
  connSock = accept( listenSock, (struct sockaddr *)NULL, (int *)NULL );
  fp = fopen("textoweb.txt","rb");
  do
  {
    num_bytes=fread( (void *)buffer, 1,1024, fp);
    ret = sctp_sendmsg( connSock, (void *)buffer, (size_t)strlen(buffer),NULL, 0, 0, 0, STREAM1, 0, 0 );
  }while(!feof(fp));
  fclose(fp);
  fp = fopen("vaquero.jpg","rb");
  do
  {
    num_bytes=fread( (void *)buffer, 1,1024, fp);
    ret = sctp_sendmsg( connSock, (void *)buffer, (size_t)strlen(buffer),NULL, 0, 0, 0, STREAM2, 0, 0 );
  }while(!feof(fp));
  fclose(fp);
  ...//Send each file with its corresponding stream identifier
}
return 0; }
```

Fig. 4. Extract of the original SCTP server code in a multistream transmission

4.1 Experimental scenario

The experimental topology is illustrated in Fig. 5. We measure both the time required to initialize the TCP or SCTP socket(s), and the time that it takes to transfer the file(s) with TCP or SCTP. Tests are carried out with two laptops in a 10 Mbps wired Ethernet local area network. Both laptops also have wireless cards to verify the multihoming feature. During the tests, there was no other traffic in the network, but the one from these experiments. Likewise, the only application running on the laptops is our TCP or SCTP application.

In the single operation tests, we transmit a 1 MB file from the server to the client through the wired local area network, and repeat the experiment for a 3MB file, and a 50MB file. Each transmission is repeated 100 times. In the multistream operation tests, the client should load a multimedia web page from the server. Therefore, the client should download a variety of multimedia files. Since we have not implemented a web server compatible with SCTP, we carry out experiments assuming that the client downloads two or four multimedia files of different sizes. Both tests (downloading two or four multimedia files) are performed 100 times. Experimental results have a confidence interval of 95% that has been calculated with a normal distribution function using 100 samples.

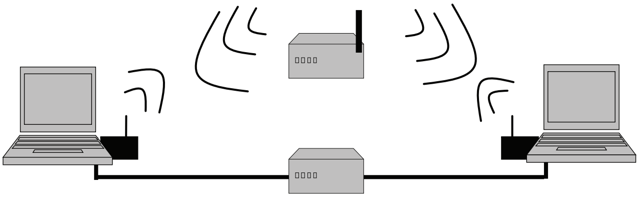


Fig. 5. Experimental topology. Laptops have Intel Centrino platforms, Intel Pentium M 740/1.73 GHz processors, and 1GB RAM. Operating system is Linux (SuSe 10.0)

4.2 Results

Results from the single operation tests show that TCP is slightly faster than SCTP in a single file transmission. Table 6 includes the average transmission time for single-file transmissions with TCP and SCTP and the corresponding confidence intervals. For instance, we observe that the transmission of a 3 MB file with SCTP lasts 2.73 seconds compared to the 2.6 seconds of TCP. SCTP is slower than TCP for two reasons. Firstly because its socket initiation time is 1ms larger (it uses four packets, adding the effect of the cookie mechanism). Secondly, the monitoring of the path that the SCTP carries out periodically (heartbeat mechanism) also introduces some overhead. As a result, the SCTP transmission lasts approximately 3% more than the TCP one.

Regarding the multistream operation, the first clear conclusion is that TCP requires more IP packets to proceed with these transmissions. A TCP connection requires three packets for negotiation and four packets for shutdown. Therefore, the more files to transmit with TCP the more packets, because it is necessary to establish a different connection to download each file (each stream) with TCP. Likewise, a SCTP association needs four packets for negotiation and three for shutdown, however, SCTP will only require an association for downloading multiple files. Fig. 6 shows the overhead amount produced with SCTP and TCP, where the x axis represents the number of files to be transmitted and the y axis the number of bytes used. We represent in this figure the number of bytes used in TCP for initiation and shutdown, as well as the number of bytes consumed by SCTP in initiation, shutdown, and heartbeat packets. For the heartbeat mechanism, we consider sending the heartbeat signal every 100ms, 250ms, 500ms, and 1 s. Observe that the time interval for sending the heartbeat is an adjustable parameter. Clearly, the more frequent the heartbeat the more bandwidth consumed. Assuming the minimum possible packet sizes for TCP and SCTP, and taking into account the SCTP heartbeat mechanism, the overhead introduced by TCP would be smaller than the one introduced by SCTP only if the heartbeat is very aggressive. Otherwise, the fact of establishing one TCP connection for each file transmission produces higher bandwidth consumption.

	1 MB file		**3 MB file**		**50 MB file**	
	TCP	**SCTP**	**TCP**	**SCTP**	**TCP**	**SCTP**
Average transmission time (s)	1.06	1.09	2.60	2.73	47.11	48.52
Confidence interval	0.24	0.29	0.32	0.20	3.78	2.26

Table 6. SCTP vs. TCP average transmission times in *single* operation tests

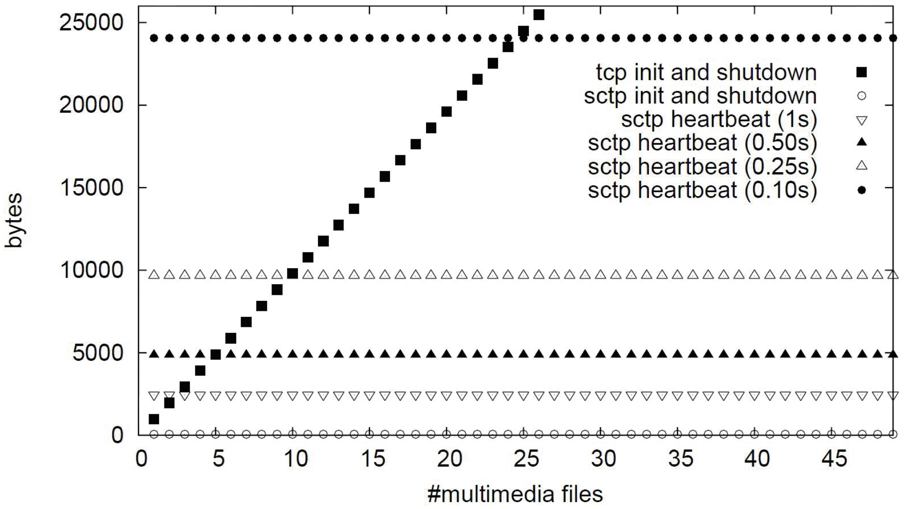


Fig. 6. Overhead introduced by TCP and SCTP. For TCP, the packet size is 20 bytes (we assume no data is sent with the first ACK packet). For SCTP, we take the following minimum packet sizes as indicated in (Stewart, 2007): INIT 20 bytes, INIT ACK 20 bytes, COOKIE ECHO 8 bytes, COOKIE ACK 4 bytes, HEARTBEAT REQUEST 4 bytes, HEARTBEAT ACK 4 bytes

On the other hand, socket initiation is still faster in TCP. However, since more sockets need to be used in TCP, the total initiation time difference between TCP and STCP is shorter and shorter as the number of files to be transmitted increases. Fig. 7 and Fig. 8 represent the duration of initiating sockets in TCP versus initiating sockets in SCTP. Indeed, when two multimedia files are transmitted (Fig. 7), the average time dedicated to sockets initiation in TCP is 1.69 ms, while the average time is 1.73 ms for SCTP. However, if we send four multimedia files, the average time increases to 3.4 ms average in TCP whereas approximately the same value remains in SCTP (Fig. 8). Thus, when four files are transmitted, SCTP total initiation time is half of the TCP total initiation time. Consequently, results show that not only the SCTP multiple file transmission is faster than the TCP one, but it consumes less bandwidth. Table 7 includes the average times for a multiple-file transmission and the corresponding confidence intervals.

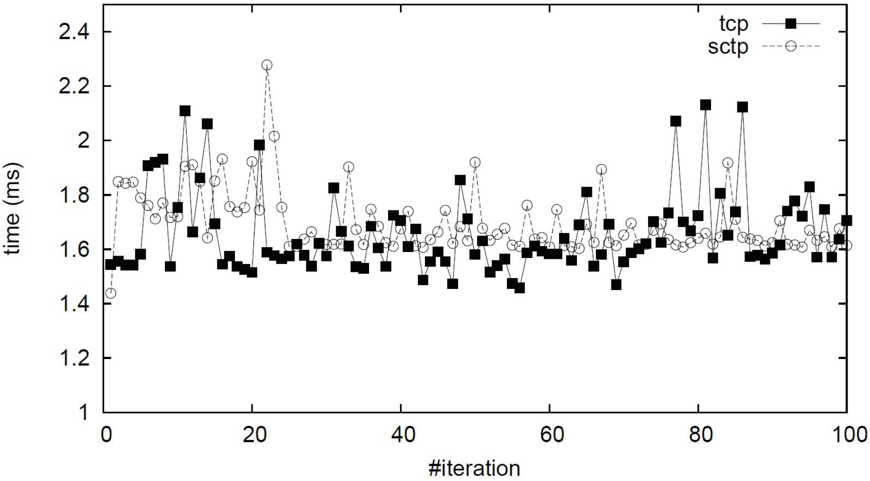


Fig. 7. Socket initiation time in TCP and SCTP in two-file downloading

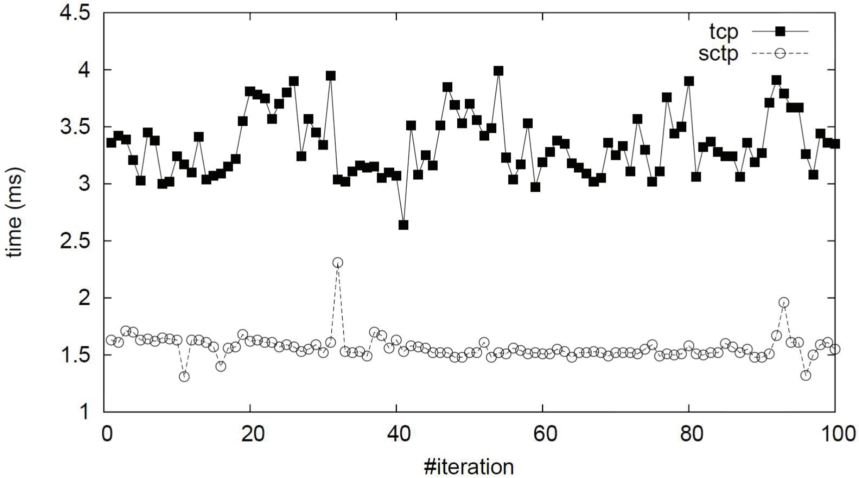


Fig. 8. Socket initiation time in TCP and SCTP in four-file downloading

	2 multimedia files		4 multimedia files	
	TCP	SCTP	TCP	SCTP
Average Transmission time (s)	6.20	3.10	18.02	6.90
Confidence intervals	1.55	0.86	1.61	1.07

Table 7. SCTP vs. TCP average transmission times in *multistream* operation tests

Finally, we test the multihoming SCTP feature in the topology shown in Fig. 5, where two PCs are connected to each other through two interfaces (one is wired, the other is wireless). We use the multistreaming SCTP client and server implementations shown in Fig. 3 and Fig. 4 respectively, including the new lines shown in Table 5. At first, client and server are using the wired network (primary IP addresses). Then, one of the wired network interface card is disabled. Experimental results show that in less than 1 second SCTP reacts in the presence of the network failure by replacing primary IP addresses with the alternative one (wireless one) to continue with the transmission. The time to change the IP addresses in use includes the ARP resolution, which is almost negligible in this scenario. Table 8 shows the exchange of IP addresses in use.

No.	Time	Source	Destination	Protocol	Info
55559	185.79019	192.168.1.10	192.168.1.11	SCTP	DATA
55560	185.79022	192.168.1.11	192.168.1.10	SCTP	SACK
55561	185.79108	192.168.1.10	192.168.1.11	SCTP	DATA
55562	185.79215	192.168.1.10	192.168.1.11	SCTP	DATA
55563	185.79218	192.168.1.11	192.168.1.10	SCTP	SACK
55564	185.79304	192.168.1.10	192.168.1.11	SCTP	DATA
55565	185.99060	192.168.1.11	192.168.1.10	SCTP	SACK
55570	186.79958	linuxpedro.local		ARP	who has 192.168.2.33? Tell 192.168.2.34
55571	186.79959	192.168.2.33		ARP	192.168.2.33 is at 00:80:5a:32:cb:c0
55572	186.80009	linuxpedro.local		ARP	who has 192.168.2.33? Tell 192.168.2.34
55573	186.80009	192.168.2.33		ARP	192.168.2.33 is at 00:80:5a:32:cb:c0
55574	186.81128	192.168.2.34	192.168.2.33	SCTP	DATA
55575	186.81132	192.168.2.33	192.168.2.34	SCTP	SACK
55576	186.83170	192.168.2.34	192.168.2.33	SCTP	DATA

```
⊕ Frame 55565 (64 bytes on wire, 64 bytes captured)
⊕ Linux cooked capture
⊕ Internet Protocol, Src: 192.168.1.11 (192.168.1.11), Dst: 192.168.1.10 (192.168.1.10)
⊕ Stream Control Transmission Protocol, Src Port: 5200 (5200), Dst Port: 20000 (20000)
        Source port: 5200
        Destination port: 20000
        Verification tag: 0x 52c9c5b0
        Checksum: 0xe5b04b29 [correct CRC32C]
        ⊕ SACK chunk (Cumulative TSN: 261016901, a_rwnd: 112640, gaps:0, TSNs: 0)
```

Table 8. Extract of the traffic captured with Wireshark (Wireshark, 2011). The first 6 SCTP packets use the primary IP addresses. After the network failure (packet# 55565), alternative addresses are used

5. Conclusion

In this work, we have presented a survey with the most relevant works on the applicability of SCTP in wireless networks. We have categorized the benefits of SCTP for wireless technologies in the following categories: mobility and handovers, multimedia transmission, and other improvements related to multiple path transmission or security. We have also shown the practical aspects of the design of a SCTP client/server application. In our example, the SCTP application is used to download files from a server. We have described the basics of how to enable multihoming and multistreaming capabilities in SCTP. We have observed that it is quite easy to adapt current applications to the SCTP protocol. When comparing to TCP, the advantages of SCTP are numerous (e.g., faster average transmission times and resources saving), above all in applications that require the transmission of multiple files. Moreover, multihoming allows increasing reliability, a key additional requirement in multimedia applications over wireless networks.

6. Acknowledgment

This research has been supported by the MICINN/FEDER project grant TEC2010-21405-C02-02/TCM (CALM).

7. References

Afif, M., Martins, P., Tabbane, S., & Godlewski, P. (2006a). Radio aware SCTP extension for handover data in EGPRS. *Proceedings 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC'06*, pp. 1-5.

Afif, M., Martins, P., Tabbane, S., & Godlewski, P. (2006b). SCTP Extension for EGPRS/WLAN Handover Data. *Proceedings 31st IEEE Conference on Local Computer Networks*, pp. 746-750.

Aydin, I., & Shen, C.-C. (2009). Performance Evaluation of Concurrent Multipath Transfer Using SCTP Multihoming in Multihop Wireless Networks. *Proceedings 8th IEEE International Symposium on Network Computing and Applications*, pp. 234-241.

Balk, A., Sigler, M., Gerla, M., & Sandidi, M. Y. (2002). Investigation of MPEG-4 video streaming over SCTP. *Proceedings 6th World Multiconference on Systemics. Cybernetics. and Informatics SCI'02*, pp. 1-4.

Begg, C. L., Pawlikowski, K., Sirisena, H., & De Silva, P. (2007). Suitability of SCTP for High Quality Video Streaming over CDMA2000. Proceedings Australasian Telecommunication Networks and Applications Conference, pp. 496-502.

Bokor, L., Huszák, A., & Jeney, G. (2009). On SCTP Multihoming Performance in Native IPv6 UMTS–WLAN Environments. *Proceedings 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops TridentCom'09*, pp. 1-10.

Cano, M.-D., Romero, J.A., & Cerdan, F. (2008). Experimental Tests on SCTP over IPSec. *Proceedings IFIP International Conference on Network and Parallel Computing NPC'08*, pp. 96-102.

Chang, L.-H., Huang, P.-H., Chu, H.-C., & Tsai, H.-H. (2009). Mobility Management of VoIP services using SCTP Handoff Mechanism. *Proceedings Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, pp. 330-335.

Cheng, R.-S., D.-J., Chao, H.-C., & Chen, W.-E. (2010). An Adaptive Bandwidth Estimation Mechanism for SCTP over Wireless Networks. *Proceedings 5th International Conference on Future Information Technology*, pp. 1-5.

Chughtai, H. M. O., Malik, S. A., & Yousaf, M. (2009) Performance Evaluation of Transport Layer Protocols for Video Traffic over WiMax. *Proceedings IEEE 13th International Multioptic Conference*, pp. 1-6.

Cui, X., Cui, L., & Koh, S. J. (2007). A Hierarchical Checksum Scheme for SCTP over Wireless Networks with Worse Channel Condition. *Proceedings International Conference on Wireless Communications, Networking and Mobile Computing WiCom'07*, pp.1845-1848.

Darche, D., Kopp, R., Mazieres, B., Lepage, F., & Gnaedinger, E. (2006). Using SCTP to improve performances of hybrid broadcast/telecommunication network system. *Proceedings of IEEE Consumer Communications & Networking Conference*, Vol. I, pp. 371-375.

Fallon, E., Murphy, L., & Murphy, J. (2009). Optimizing Metropolitan Area Wireless Path Selection Using Media Independent Handover. *Proceedings Second International Workshop on Cross Layer Design IWCLD'09*, pp. 1-5.

Honda, M., Sakakibara, H., Nishida, Y., & Tokuda, H. (2007). SmSCTP: A Fast Transport Layer Handover Method Using Single Wireless Interface. *Proceedings 12th IEEE International Symposium on Computers and Communications ISCC'07*, pp.319-324.

Huang, C.-M., & Lin, M.-S. (2010). RG-SCTP: Using the Relay Gateway Approach for Applying SCTP in Vehicular Networks. *Proceedings IEEE International Symposium on Computers and Communications ISCC'10*, pp. 139-144.

IEEE Std 802.11-1997. (1997). IEEE 802.11 wireless LAN medium access control (MAC) and physical layer (PHY) specifications.

IEEE Std. 802.16e-2005. (2005). IEEE Standard for Local and metropolitan area networks. Part 16: Air interface for fixed broadband wireless access systems.

IEEE Std. 802.21-2008. (2008). IEEE Standard for Local and metropolitan area networks. Part 21: Media Independent Handover Services.

Iyengar, J. R., Amer, P., & Stewart, R. (2006). Concurrent multipath transfer using SCTP multihoming over independent end-to-end paths. *IEEE/ACM Transactions on Networking*, Vol. 14, No. 5, pp. 951–964.

Kamal, H., Penoff, B., & Wagner, A. (2005). SCTP versus TCP for MPI. *Proceedings of ACM/IEEE SuperComputing Conference SC'05*, pp. 30-44.

Kim, D., Song, J., Kim, J., Yoo, H., Park, J., & Cano, J.C. (2006). The Applicability of SCTP to Mobile Ad Hoc Networks. *Proceedings International Conference on Advanced Communication Technology ICACT'06*, Vol. 3, pp. 1979-1984.

Kohler, E., Handley, M., & Floyd, S. (2006). Datagram Congestion Control Protocol. RFC 4340.

Kozlovszky, M., Berceli, T., & Kutor, L. (2006). Analysis of SCTP and TCP based communications in high speed clusters. *Nuclear Instruments and Methods in Physics Research Section A*, Vol. 559, Issue 1, pp.85-896.

Lee, Y.-J. & Atiquzzaman, M. (2009). Mean Waiting Delay for Web Object Transfer in Wireless SCTP Environment. *Proceedings IEEE International Conference on Communications ICC'09*, pp. 1-5.

Lee, Y.-J., Lee, D.-W., & Atiquzzaman, M. (2009). Novel web agent framework to support seamless mobility for data networks. *IET Communications Journal*, Vol. 3, No. 12, pp. 1861–1869.

Leu, F.-Y. & Ko, Z.-J. (2008). A Novel Network Mobility Scheme Using SIP and SCTP for Multimedia Applications. *Proceedings of International Conference on Multimedia and Ubiquitous Engineering*, pp. 564-569.

Liu, H.-S., Hsieh, C.-C., Chen, H.-C., Hsieh, C.-H., Liao, W., Chu, P.-C., & Wang, C.-H. (2010). Exploiting Multi-link SCTP for Live TV Broadcasting Service. *Proceedings IEEE 71st Vehicular Technology Conference*, pp. 1-6.

Ma, L., Yu, F., Leung, V. C. M., & Randhawa, T. (2004). A new method to support UMTS/WLAN vertical handover using SCTP. IEEE Wireless Communications, Vol. 11, No. 4, pp. 44-51.

Ma, L., Yu, F. R., & Leung, V. V. M. (2007). Performance Improvements of Mobile SCTP in Integrated Heterogeneous Wireless Networks. *IEEE Transactions on Wireless Communications*, Vol. 6, No. 10, pp. 3567-3577.

Mascolo, S., Grieco, L. A., Ferorelli, R., Camarda, P., & Piscitelli, G. (2004). Performance evaluation of Westwood+ TCP congestion control. *Performance Evaluation*, Vol. 4, No. 55, pp. 93–111.

Natarajan, P., Iyengar, J. R., Amer, P. D., & Stewart, R. (2006). SCTP: An innovative transport layer protocol for the web. *Proceedings 15th International World Wide Web Conference, WWW'06*, pp. 615-624.

Nosheen, S., Malik, S. A., Zikria, Y. B., & Afzal, M., K. (2007). Performance Evaluation of DCCP and SCTP for MPEG4 Video over Wireless Networks. *Proceedings IEEE 11th International Multitopic Conference*, pp. 1-6.

Perotto, F., Casetti, C., & Galante, G. (2007). SCTP-based Transport Protocols for Concurrent Multipath Transfer. *Proceedings IEEE Wireless Communications and Networking Conference WCNC'07*, pp. 2969-2974.

Shaojian, F., Atiquzzaman, M., & Ivancic, W. (2005). Evaluation of SCTP for space networks. *IEEE Wireless Communications*, Vol. 12, No. 5, pp. 54-62.

Shieh, C.-S., Lin, I-C., & Lai, W. K. (2008). Improvement of SCTP Performance in Vertical Handover. *Proceedings of Eighth International Conference on Intelligent Systems Design and Applications*, pp. 494-498.

Stewart, R. (2007). Stream Control Transmission Protocol. RFC 4960.

Stewart, R., Xie, Q., Tuexen, M., Maruyama, S., & Kozuka, M. (2007). Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration. RFC 5061.

Wang, B., Feng, W., Zhang, S.-D., & Zhang, H.-K. (2010). Concurrent multipath transfer protocol used in ad hoc networks. *IET Communications Journal*, Vol. 4, No. 7, pp. 884–893.

Wang, H., Jin, Y., Wang, W., Ma, J., & Zhang, D. (2003). The performance comparison of PRSCTP, TCP and UDP for MPEG-4 multimedia traffic in mobile network. *Proceedings International Conference on Communication Technology (ICCT),* Vol. 1, pp. 403-406.

Wang, L., Kawanishi, K., & Onozato, Y. (2008). MPEG-4 Optimal Transmission over SCTP Multi-streaming in 802.11 Wireless Access. *Proceedings 7th Asian-Pacific Symposium on Information and Telecommunication Technologies*, pp. 172-177.

Wang, L., Kawanishi, K., & Onozato, Y. (2009).Achieving Robust Fairness of SCTP Extension for MPEG-4 Streaming. *Proceedings 20th Personal, Indoor and Mobile Radio Communications Symposium PIMRC '09*, pp. 2970-2974.

Wireshark. <http://www.wireshark.org>. Last visited March 20th, 2011.

Xu, C., Fallon, E., Qiao, Y., Zhong, L., & Muntean, G.-M. (2011). Performance Evaluation of Multimedia Content Distribution Over Multi-Homed Wireless Networks. *IEEE Transactions on Broadcasting*, Vol. PP (99), pp. 1-12.

Yuan, Y., Zhang, Z., Li, J., Shi, J., Zhou, J., Fang, G., & Dutkiewicz, E. (2010). Extension of SCTP for Concurrent Multi-Path Transfer with Parallel Subflows. *Proceedings IEEE Wireless Communications and Networking Conference WCNC'10*, pp. 1-6.

13

Traffic Control for Composite Wireless Access Route of IEEE802.11/16 Links

Yasuhisa Takizawa
Kansai University
Japan

1. Introduction

The expansion and diversification of wireless communications are proceeding rapidly with the diffusion of cellular phones, WiFi and WiMAX. However, concern is increasing that the growth of wireless systems will exhaust finite wireless resources. Cognitive radio technology(Mitorall & Maguire, 1999; Mitoralll, 1999; Harada, 2005), which has been proposed as a solution to this problem, aims to optimize the utilization of diverse wireless resources. Furthermore, AIPN (All-IP Network) (3GPP, 2005) and NGN (Next Generation Network)(ITU, 2006) investigate the network architecture that accommodates diverse communication media. Accordingly, we expect that in the near future, wireless access networks will be composed of diverse wireless medias.

To exploit wireless media diversity in expected access networks, some bandwidth-aggregation methods in wireless media have recently been proposed. Bandwidth-aggregation combines diverse communication links in parallel and suitably distributes packets to communication links. The works(Phatak & Goff, 2002; Snoeren, 1999; Shrama et al., 2007) aggregate wireless links in IP to improve IP throughput. The work(Chebrou & Rao, 2006) also aggregates wireless links in IP to decrease IP delay based on wireless media that provide a bandwidth guarantee. The works(Hsieh et al., 2004; Zhang et al., 2004) aggregate communication links in a transport layer to improve TCP throughput. Meanwhile, wireless access networks process traffic of diverse application, and the traffic is classified by the following two types of application traffic:

- Traffic of throughput-oriented application such as FTP and Web on TCP.
- Traffic of delay-oriented application such as VoIP and Video Conference on UDP.

Therefore, wireless access networks are required to provide high throughput and low delay by diverse applications. The above works do not consider delay except for the work(Chebrou & Rao, 2006), and the work(Chebrou & Rao, 2006) does not consider IEEE802.11 that no bandwidth guarantee is provided. Furthermore, the works(Phatak & Goff, 2002; Snoeren, 1999; Shrama et al., 2007; Chebrou & Rao, 2006) improve IP performance, but can not provide effective improvement of application performance because they do not consider out-of-order packets which occur by the packet distribution to multiple links. The works(Hsieh et al., 2004; Zhang et al., 2004) consider the out-of-order packet, and can improve the performance of TCP application, but can not improve that of UDP application such as VoIP and Video Conference.

	802.11a/b	802.16
Transmission Rate	54Mbps/11a, 11Mbps/11b	75Mbps
Coverage	50m/11a, 100m/11b	1000m
Access Control	CSMA/CA (Decenteralized)	TDD/FDD (Centeralized)
Bandwidth Guarantee	No	Yes

Table 1. Performance of wireless systems.

In this chapter, assuming the expected wireless access network to be composed of IEEE802.11, which is a popular wireless system, and IEEE802.16, which is expected to spread, a IP packet distribution on the access route, which combines IEEE802.11-link and IEEE802.16-link in parallel, is proposed to improve the application performance. The proposed packet distribution increases IP throughput and decreases IP delay. Furthermore, it reduces out-of-order packets and provides high throughput and low delay to both UDP applications and TCP applications simultaneously.

Our works(Takizawa et al., 2008; Takizawa, 2008) have proposed the packet distribution for combining IEEE802.11/16 wireless upload links. We expand the above packet distribution to reduce out-of-order packets and to apply download traffic, and show its essential characteristics of packet distribution for composite wireless access route of IEEE802.11/16-links (call M-route) , then propose a packet distribution method for M-route. Furthermore, we evaluate the method's performance by multiple application traffic on both UDP and TCP in a wireless access network composed of 802.11a, 802.11b and 802.16, which have the different characteristic from each other (see Table 1).

The configuration of wireless access networks by wireless media diversity is assumed as follows (see Fig. 1).

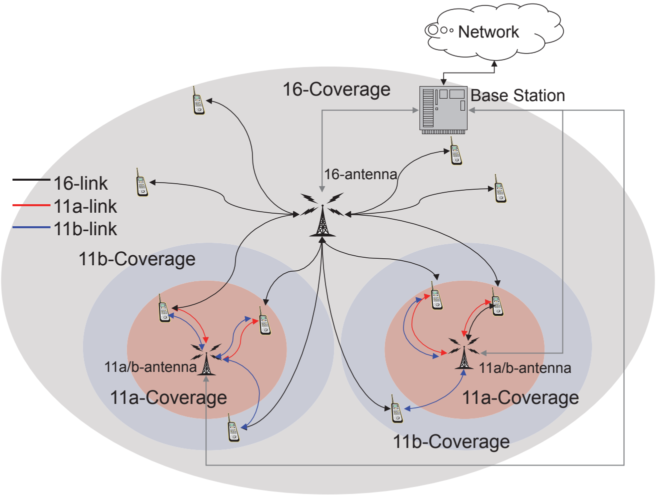


Fig. 1. Assumed wireless access network.

- Base station provides an access point function of IEEE802.11a/b-wireless systems and a base station function of 16-wireless system, and accommodates IEEE802.11a/b-antennas and an IEEE802.16-antenna by wired connecting. It also provides the function of gateway.

- Each terminal is equipped with IEEE802.11a/b-interfaces and IEEE802.16-interface, and can communicate with base station by using each interface.
- IEEE802.11a/b-antennas and terminals are randomly deployed within coverage of IEEE802.16-antenna.
- The access network is IP network.

2. Characteristics of IEEE802.11 link for packet distribution

In this section, based on Media Access Control (MAC) of IEEE802.11 DCF, the characteristics of IEEE802.11 wireless link (11-link) for packet distribution is analyzed.

2.1 IEEE802.11 link cost

Based on queuing theory(Gross & Harris, 1985), a link load is shown as the number of packets in a link, including waiting packets in the queue and the currently processed packet. $d_{(i,k)}$, which is cost of link k between a terminal i and a base station, is defined as the link load, and it is expressed using Little's theorem(Little, 1961) as follows.

$$d_{(i,k)} = F_{(i,k)} \cdot T_{(i,k)} \tag{1}$$

where $F_{(i,k)}$ is the packet arrival rate of link k in terminal i and $T_{(i,k)}$ is the average delay of link k in terminal i. Delay is the time from packet arrival at the terminal to completion of packet transmission, therefore the delay is composed of a waiting delay in queue and an air time. The air time is composed of MAC delay and transmission delay, which take the MAC retransmission into consideration.

Based on Eq. (1), $T_{(i,k)}$ decreases if $d_{(i,k)}$ decreases on constant $F_{(i,k)}$ and on maximum of $d_{(i,k)}$, that is, link capacity, $F_{(i,k)}$ can increase if $T_{(i,k)}$ decreases. $F_{(i,k)}$ corresponds to a throughput on condition that no packet loses. Therefore, when $d_{(i,k)}$ decreases, a throughput increases and a delay decreases on a link.

The dependence of the link cost on the packet arrival rate, which corresponds to the number of distributed packets in unit time to a link, is shown. Based on Eq. (1), the link cost depends on the average delay. The average delay is composed of the waiting delay in queue and the packet service time. Therefore, in regard with 11-link, the dependence of the above elements on the packet arrival rate are shown, and in summarizing them, the dependence of the link cost on the packet arrival rate is shown.

2.1.1 Dependence of packet service time on packet arrival rate

In (Bianchi, 2000), throughput analysis of IEEE802.11 DCF is shown, and in (Carvalho & Garcia, 2003), the packet service time analysis of that is shown based on (Bianchi, 2000). According to these, the dependence of the average packet service time on the packet arrival rate is shown.

DCF adopts an exponential backoff scheme, and employs a discrete-time backoff timer. The timer immediately following a Distributed InterFrame Space (DIFS) starts, and a terminal, which is a terminal or a base station, is allowed to transmit only at the beginning of each Slot Time. The Slot Time size σ is set equal to the time needed at any terminal to detect the transmission of a packet from any other terminal. At each packet transmission, the backoff timer is randomly chosen in the range $(0, CW - 1)$. CW is called Contention Window, and depends on the number of transmissions failed for the packet. At the first transmission attempt, CW is set equal to CW_{min} called minimum contention window. After each failed transmission, CW is doubled, up to a maximum value $CW_{max} = 2^r CW_{min}$ (r is a maximum

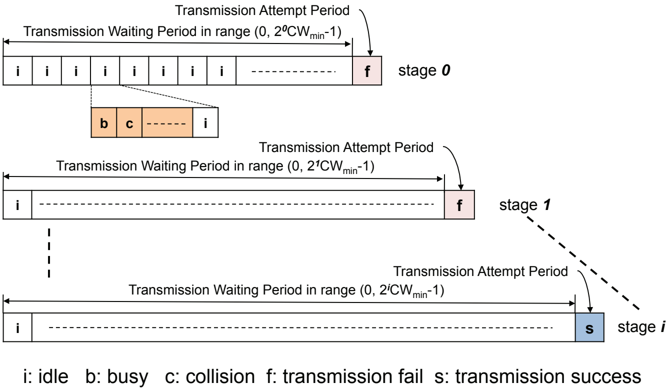

Transmission Attempt Period
Transmission Waiting Period in range (0, 2⁰CWmin-1)
i
b
c
f
stage 0
Transmission Attempt Period
Transmission Waiting Period in range (0, 2¹CWmin-1)
stage 1
Transmission Attempt Period
Transmission Waiting Period in range (0, 2ⁱCWmin-1)
s
stage i
i: idle b: busy c: collision f: transmission fail s: transmission success

Fig. 2. Exponential binary backoff in IEEE802.11.

number of retransmissions). Each transmission attempt is referred to as a bakoff stage. The packet service time is the sum of time for each backoff stage. Each backoff stage is composed of the transmission waiting period and the transmission attempt period (see Fig. 2). The backoff stage starts in the transmission waiting period, and the backoff timer is initialized to a random value in the range $(0, CW_i - 1)$ at the backoff stage i start. CW_i is the contention window size of the backoff stage i. In the period, the backoff timer is decremented only when the channel is idle, and it is frozen when the channel is busy. The duration of the period is the time until the backoff timer becomes zero from initial value. The transmission attempt period starts when the backoff timer reaches zero, and a packet transmission takes place. The duration of period is the time to transmit a packet. In the model of (Bianchi, 2000) and (Carvalho & Garcia, 2003), a fixed number of terminals is assumed, and the backoff stage is repeated until a packet transmission success using CW_i until stage r and using CW_r beyond stage r. The stage r is called maximum backoff stage. Furthermore, using the probability τ that a terminal transmits in a randomly chosen slot time, the following probabilities in an exponential backoff scheme are expressed.

$$\begin{aligned}
p_{tr} &= 1 - (1-\tau)^{n-1} \\
p_{suc} &= \frac{(n-1)\tau(1-\tau)^{n-2}}{p_{tr}} \\
p_i &= 1 - p_{tr} \\
p_s &= p_{tr} \cdot p_{suc} \\
p_c &= p_{tr}(1 - p_{suc}) \\
q &= (1-\tau)^{n-1}
\end{aligned} \tag{2}$$

where n is the number of terminal in the channel coverage, p_{tr} is the probability that there is at least one transmission in the slot time of the transmission waiting period, p_{suc} is the probability that a transmission occurring on the channel is successful, p_i is the probability that the slot time is idle in the transmission waiting period, p_s is the probability that the channel is busy due to a packet transmission success in the transmission waiting period, p_c is the probability that the channel is busy due to a collision in the transmission waiting period, and q is the probability that a packet transmission success in the transmission attempt period. Let B be the average time which the transmission waiting period takes until a packet transmission succeeds, and let A be the average time which the transmission attempt period takes until a

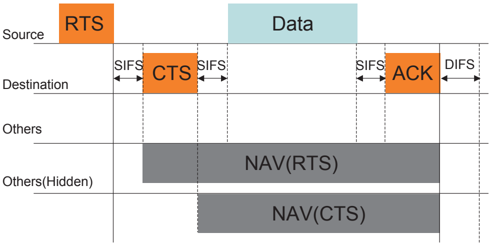


Fig. 3. RTS/CTS access control sequence in IEEE802.11.

packet transmission succeeds, B and A are derived from a binary exponential backoff scheme as follows(Carvalho & Garcia, 2003). (Note: In this section, "time" is the duration in slot time units σ of IEEE802.11)

$$B = \frac{t_b(\eta CW_{min} - 1)}{2q} \tag{3}$$

$$\begin{aligned} t_b &= p_i t_i + p_s t_s + p_c t_c \\ \eta &= \frac{q - 2^r(1-q)^{r+1}}{1 - 2(1-q)} \end{aligned} \tag{4}$$

$$A = \frac{1-q}{q} t_c + t_s \tag{5}$$

$$\begin{aligned} t_i &= 1 \\ t_s &= RTS + SIFS + \delta + CTS + SIFS + \delta + H \\ &\quad + PL + SIFS + \delta + ACK + DIFS + \delta \\ t_c &= RTS + DIFS + \delta \end{aligned} \tag{6}$$

where t_i is the time of idle (i.e., one backoff slot), t_s is the average time that the channel is sensed busy due to a packet transmission success, t_c is the average time that the channel is busy due to a collision in the channel, RTS, CTS and ACK are time that RTS, CTS and ACK frame is transmitted respectively, $SIFS$ and $DIFS$ are the interval time (see Fig. 3), δ is the propagation delay, H is the time that a packet header is transmitted, and PL is the time the payload is transmitted. According to Eq. (2), $q = 1 - p_{tr}$, therefore, t_b/q expresses the average time that the backoff timer is decreased by one, and $(\eta CW_{min} - 1)/2$ expresses the average of sum of backoff timer in all stage. In Eq. (5), $(1-q)/q$ expresses the average number of collision in the transmission attempt priod.

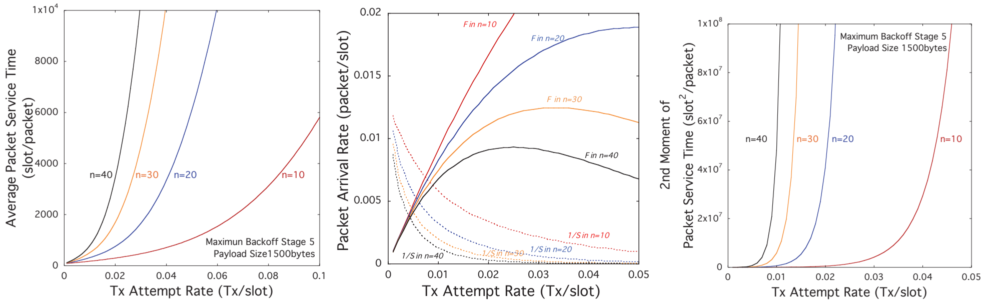
Then, the average packet service time S is argued using the above analysis. S is shown as follows.

$$S = B + A \tag{7}$$

When the number of terminal is constant, the dependence of S on τ is shown using the first and second derivative of S at τ as follows.

$$\frac{dS}{d\tau} > 0 \qquad \frac{d^2S}{d\tau^2} > 0 \tag{8}$$

Therefore, S is a convex monotonically increasing function of τ. Figure 4(a) illustrates the dependence of S on τ by using Eq. (7) in 11b MAC parameter, transmission rate 11Mbps, a number of terminals 10∼40, maximum backoff stage 5, and the payload size 1500 bytes, and it also shows the same characteristics.



(a) Dependence of S on τ.

(b) Dependence of F on τ.

(c) Dependence of $\overline{S^2}$ on τ.

Fig. 4. Dependence of each element on τ.

In (Bianchi, 2000) and (Carvalho & Garcia, 2003), the transmission queue is assumed to be always non-empty, thus, the dependence of τ on the packet arrival rate F is not considered. Let F be the number of arrival packets at a link in a slot time, the dependence is argued. The average number of arrival packets in period S is FS, and the average number of transmission attempts on a successfully transmitted packet is $(1-q)/q+1$. Then, the average number of that a packet transmission attempts in period S is FS/q. Therefore, τ is shown as follows.

$$\tau = \frac{FS}{qS} = \frac{F}{q} \tag{9}$$

Figure 4(b), which illustrates the dependence of F on τ using Eq. (9) in the same parameter as Fig. 4(a). In Fig. 10, when $F < 1/S$ ($1/S$ is the packet service rate), that is, when the load does not exceed the link capacity, and when the number of terminal is constant, F for τ is concavely and monotonically increasing. Therefore, within link capacity, the dependence of F on τ is shown using the first and second derivative of F at τ as follows.

$$\frac{\mathrm{d}F}{\mathrm{d}\tau} > 0 \qquad \frac{\mathrm{d}^2F}{\mathrm{d}\tau^2} < 0 \tag{10}$$

Furthermore, the first and second derivative of S on F is shown using Eqs.(8), (10) as follows.

$$\frac{\mathrm{d}S}{\mathrm{d}F} = \frac{\mathrm{d}S}{\mathrm{d}\tau}\frac{\mathrm{d}\tau}{\mathrm{d}F} = \frac{\mathrm{d}S}{\mathrm{d}\tau}\frac{1}{\left(\dfrac{\mathrm{d}F}{\mathrm{d}\tau}\right)} > 0 \tag{11}$$

$$\frac{\mathrm{d}^2S}{\mathrm{d}F^2} = \frac{\mathrm{d}^2S}{\mathrm{d}\tau^2}\left(\frac{\mathrm{d}\tau}{\mathrm{d}F}\right)^2 - \frac{\mathrm{d}S}{\mathrm{d}\tau}\frac{1}{\left(\dfrac{\mathrm{d}^2F}{\mathrm{d}\tau^2}\right)} > 0 \tag{12}$$

Therefore, within link capacity, S is a convex monotonically increasing function of F.

2.1.2 Dependence of waiting delay in queue on packet arrival rate

The dependence of W which is the waiting delay in queue on F is argued. N_Q, which is the number of waiting packets in queue, is $F \times W$ using Little's theorem. W is composed of the packet service time for N_Q packets and R, which is the sum of the residual service time in each packet arrival. Consequently, W is shown as follows.

$$W = N_Q \cdot S + R = F \cdot W \cdot S + R \tag{13}$$

Each residual service time in a packet arrival is $\overline{S^2}/2S$(Bertsetkas & Gallager, 1992), where $\overline{S^2}$ is the second moment of S. The average number of packet arrivals in S is FS; accordingly, R is $F\overline{S^2}/2$. Applying the above relations to Eq. (13), W is given as

$$W = \frac{F\overline{S^2}}{2(1 - FS)} \tag{14}$$

Let $V[S]$ be the variance of S, and it is shown as follows(Carvalho & Garcia, 2003)

$$\begin{aligned} V[S] &= \left[\frac{t_b(CW_{min}\gamma - 1)}{2} + t_c\right]^2 \frac{1-q}{q^2} \\ \gamma &= \frac{[2q^2 - 4q + 1 - r(-1 + 2q)q][2(1-q)]^r + 2q^2}{(-1+2q)^2} \end{aligned} \tag{15}$$

Using Eq. (15), $\overline{S^2}$ is shown as follows.

$$\overline{S^2} = S + V(S) \tag{16}$$

Furthermore, using Eq. (16), the first and second derivatives of $\overline{S^2}$ at τ are shown, respectively, as follows.

$$\frac{\mathrm{d}\overline{S^2}}{\mathrm{d}\tau} > 0 \quad \frac{\mathrm{d}^2\overline{S^2}}{\mathrm{d}\tau^2} > 0 \tag{17}$$

Figure 4(c) illustrates the dependence of $\overline{S^2}$ on τ using Eq. (16) in the same parameter as Fig. 4(a), and it also shows the same characteristics. Furthermore, applying Eq. (10) to Eq. (17), the first and second derivatives of $\overline{S^2}$ at F are shown, respectively, as follows.

$$\frac{\mathrm{d}\overline{S^2}}{\mathrm{d}F} > 0 \quad \frac{\mathrm{d}^2\overline{S^2}}{\mathrm{d}F^2} > 0 \tag{18}$$

Using Eqs. (14) (18), the first and second derivatives of W at F are shown, respectively, on the condition of $FS < 1$, as follows.

$$\frac{\mathrm{d}W}{\mathrm{d}F} > 0 \quad \frac{\mathrm{d}^2W}{\mathrm{d}F^2} > 0 \tag{19}$$

$FS < 1$, that is, $F < 1/S$ expresses the condition that a link load is with a link capacity. Therefore, within a link capacity, W is also a convex monotonic increasing function of F.

2.1.3 Dependence of 11-link cost on packet arrival rate

Finally, the dependence of the 11-link cost on the packet arrival is argued. The average delay T is also a convex monotonic increasing function of F because of $T = W + S$. Applying the dependence of T on F to Eq. (1), the first and second derivatives of a 11-link cost d at F are as follows.

$$\frac{\mathrm{d}d}{\mathrm{d}F} > 0 \quad \frac{\mathrm{d}^2 d}{\mathrm{d}F^2} > 0 \tag{20}$$

Consequently, a 11-link cost d is also a convex monotonic increasing function of F within a link capacity and in a fixed number of terminals.

2.2 Cost of M-route compositing multiple 11-links for upload traffic

On communications using a M-route which aggregates multiple 11-links from terminal to a base station , the cost of M-route for upload traffic is the sum of cost of each 11-uplink composing M-route because the number of packets in a M-route is the sum of the number of packets in each link composing M-route. Therefore, m_i which is the cost of M-route for upload traffic in terminal i is shown as follows (see Fig. 5(a)).

$$m_i = \sum_{x \in U_i} d_{(i,x)} \tag{21}$$

U_i is the set of an uplink which is provided by a 11-wireless interface equipped with terminal i. Here, in steady packet arrival rate, the packet distribution from an 11-uplink k to an 11-uplink j in M-route of terminal i, is argued. In this case, the packet distribution to the other 11-uplinks is constant, thus the dependence of $F_{(i,j)}$ on $F_{(i,k)}$ is shown as follows.

$$\frac{\mathrm{d}F_{(i,j)}}{\mathrm{d}F_{(i,k)}} = -1 \quad \frac{\mathrm{d}^2 F_{(i,j)}}{\mathrm{d}(F_{(i,k)})^2} = 0 \tag{22}$$

Using Eqs.(20) and (22), the first and second derivatives of $d_{(i,j)}$ at $F_{(i,k)}$ are shown as follows.

$$\begin{aligned} \frac{\mathrm{d}d_{(i,j)}}{\mathrm{d}F_{(i,k)}} &= \frac{\mathrm{d}d_{(i,j)}}{\mathrm{d}F_{(i,j)}} \frac{\mathrm{d}F_{(i,j)}}{\mathrm{d}F_{(i,k)}} = -\frac{\mathrm{d}d_{(i,j)}}{\mathrm{d}F_{(i,j)}} < 0 \\ \frac{\mathrm{d}^2 d_{(i,j)}}{\mathrm{d}(F_{(i,k)})^2} &= \frac{\mathrm{d}^2 d_{(i,j)}}{\mathrm{d}(F_{(i,j)})^2} \left(\frac{\mathrm{d}F_{(i,j)}}{\mathrm{d}F_{(i,k)}} \right)^2 + \frac{\mathrm{d}d_{(i,j)}}{\mathrm{d}F_{(i,j)}} \frac{\mathrm{d}^2 F_{(i,j)}}{\mathrm{d}(F_{(i,k)})^2} > 0 \end{aligned} \tag{23}$$

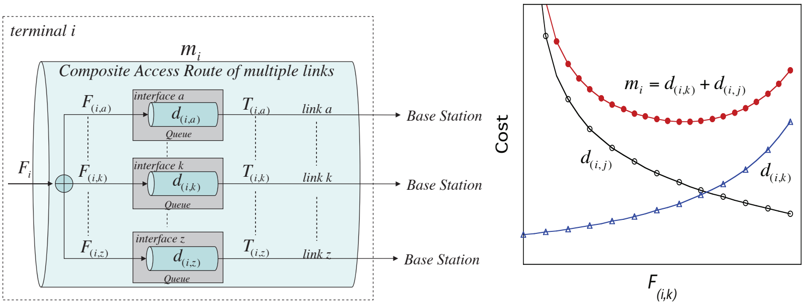
Consequently, $d_{(i,j)}$ is a convex monotonically decreasing function of $F_{(i,k)}$. According to Eq.(21), m_i is the sum of $d_{(i,k)}$, which is a convex monotonically increasing function of $F_{(i,k)}$, and $d_{(i,j)}$, which is a convex monotonically decreasing function of $F_{(i,k)}$, and the uplink cost of the others, which are constant for $F_{(i,k)}$. Therefore, m_i is a convex function of $F_{(i,k)}$ (see Fig.5(b)), and m_i has a optimal solution for $F_{(i,k)}$.

Because m_i is a convex function of $F_{(i,k)}$, the optimal solution can be searched by the packet distribution which aim to descend the gradient in the convex function. When packets are distributed from a 11-uplink k to 11-uplink j in M-route, the condition of the gradient descent on M-route cost is shown as follows using Eq.(22).

$$\frac{\mathrm{d}m_i}{\mathrm{d}F_{(i,k)}} = \frac{\mathrm{d}d_{(i,k)}}{\mathrm{d}F_{(i,k)}} - \frac{\mathrm{d}d_{(i,j)}}{\mathrm{d}F_{(i,j)}} > 0 \tag{24}$$

Applying Eq.(1) to Eq.(24), and transforming Eq.(24) into difference equation, thus the first derivative of m_i at $F_{(i,k)}$ is shown as follows.

$$\frac{\mathrm{d}m_i}{\mathrm{d}F_{(i,k)}} = \lim_{\Delta F_{(i,k)} \to 0} \left(\left(T_{(i,k)} + F_{(i,k)} \frac{\Delta T_{(i,k)}}{\Delta F_{(i,k)}} \right) - \left(T_{(i,j)} + F_{(i,j)} \frac{\Delta T_{(i,j)}}{\Delta F_{(i,j)}} \right) \right) > 0 \tag{25}$$



(a) M-route cost for upload traffic.

(b) Dependence of M-route cost on packet distribution.

Fig. 5. M-route for upload traffic.

Furthermore, applying finite difference approximation to Eq.(25), the following is derived.

$$\frac{\mathrm{d}m_i(n)}{\mathrm{d}F_{(i,k)}} \approx T_{(i,k)}(n+1) - T_{(i,j)}(n+1) > 0 \tag{26}$$

Where, $m_i(n)$ is M-route cost of terminal i in packet distribution of n time and $T_{(x,y)}(n+1)$ is average delay of 11-link y in terminal x in packet distribution of $n+1$ time.
Consequently, when the packet distribution meets Eq.(26) which means the average delay of source 11-uplink on packet distribution becomes larger than that of destination 11-uplink on packet distribution, the M-route cost for upload traffic decreases and approaches the optimal solution. Such packet distribution is repeated with the decrease in the amount of the distributing packets ($\Delta F_{(i,k)} \to 0$), and finally the average delay of source 11-uplink becomes equal to that of destination 11-uplink, the M-route cost for upload traffic reaches its optimal solution.
Furthermore, the search for the optimal solution of M-route cost has the additional effectiveness which decreases the arrival of out-of-order packets because of the equalization of the delay of source 11-link and destination 11-link.

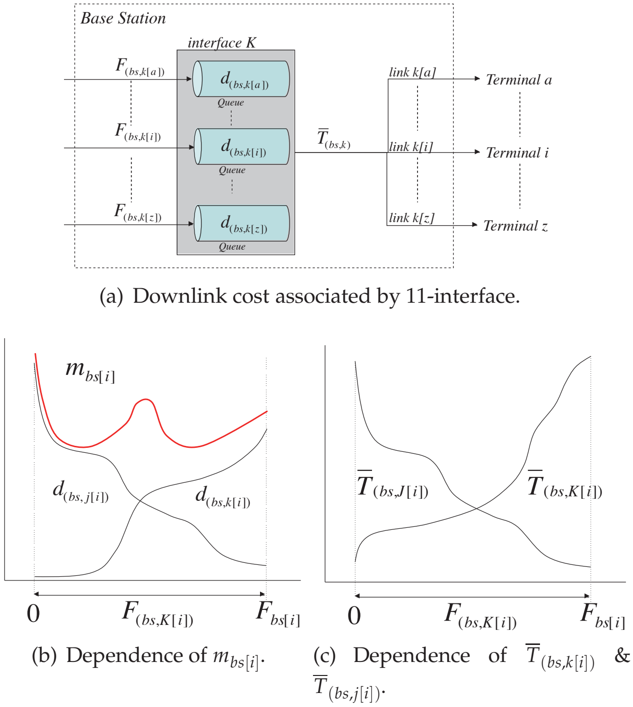
2.3 Cost of M-route compositing multiple 11-Links for download traffic

A base station associates its 11-interface with multiple terminals in its coverage. Thus, its interface is composed of multiple 11-downlinks according to multiple terminals in its 11-coverage, that is, its topology is point-to-multipoint. In this subsection, the cost of M-route for download traffic (i.e. in a base station) in steady packet arrival rate is argued.
In queueing theory, a link has a queue of packets to be transmitted, and has an independent server on other links within the same interface. However, an 11-downlink is different from a link reserved the resource such as WiMAX (TDD or FDD) link and CDMA link, and an 11-downlink shares the resource of interface among other downlinks within the same interface. Conceptually, we can also view an 11-downlink within an interface as follows.

- Each 11-downlink has a queue which is independent on the other downlinks.
- Each 11-downlink has a common server as an interface among the other downlinks.

That is, in 11-downlink $k[i]$ to terminal i, which is provide by interface k of base station, $F_{(bs,k[i])}$ which is packet arrival rate of link $k[i]$ in base station, is independent on the others, and $T_{(bs,k[i])}$ which is average delay of link $k[i]$ in base station, is common among the others. Therefore, $d_{(bs,k[i])}$ which is cost of 11-link $k[i]$ in base station, is shown as follows (see Fig. **??**).

$$d_{(bs,k[i])} = F_{(bs,k[i])} \cdot \overline{T}_{(bs,k[i])} \tag{27}$$



(a) Downlink cost associated by 11-interface.

(b) Dependence of $m_{bs[i]}$.

(c) Dependence of $\overline{T}_{(bs,k[i])}$ & $\overline{T}_{(bs,j[i])}$.

Fig. 6. Downlink and M-route for download traffic.

where $\overline{T}_{(bs,k[i])}$ is the average delay of 11-interface k providing downlink $k[i]$ in base station. That is, $\overline{T}_{(bs,k[i])}$ is the average delay based on all the packets which are distributed to 11-interface k.

To argue the dependence of $d_{(bs,k[i])}$ on $F_{(bs,k[i])}$, the first derivative of $d_{(bs,k[i])}$ at $F_{(bs,k[i])}$ is shown. Using Eq.(27), it is shown as follows.

$$\frac{\mathrm{d}d_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}} = \overline{T}_{(bs,k[i])} + F_{(bs,k[i])} \frac{\mathrm{d}\overline{T}_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}} \tag{28}$$

It is difficult to derive $\frac{\mathrm{d}\overline{T}_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}}$, which is the dependence of of $\overline{T}_{(bs,k[i])}$ on $F_{(bs,k[i])}$, because $\overline{T}_{(bs,k[i])}$ is dependent on not only $F_{(bs,k[i])}$ but also the packet distribution of the other downlinks provided by 11-interface k. To simplify this difficulty, the following condition is assumed.

$$\frac{\mathrm{d}\overline{T}_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}} > 0 \tag{29}$$

According to the condition Eq.(29), its is $\frac{\mathrm{d}d_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}} > 0$, then $d_{(bs,k[i])}$ is a monotonically increasing function of $F_{(bs,k[i])}$.

In the condition, the packet distribution from a 11-downlink $k[i]$ to a 11-downlink $j[i]$, is argued. These 11-downlinks are contained in the M-route which aggregates 11-downlinks to terminal i, and are respectively provided by different 11-interface (11-interface k and j). The same as the packet distribution of M-route for upload traffic, the packet distribution to the other 11-downlinks to terminal i, which is respectively provided by different 11-interface except for 11-interface k and j, is constant, thus the dependence of $F_{(bs,j[i])}$ on $F_{(bs,k[i])}$ is shown as follows.

$$\frac{\mathrm{d}F_{(bs,j[i])}}{\mathrm{d}F_{(bs,k[i])}} = -1$$
$$\frac{\mathrm{d}^2 F_{(bs,j[i])}}{\mathrm{d}(F_{(bs,k[i])})^2} = 0 \tag{30}$$

Therefore, the first derivative of $d_{(bs,j[i])}$ at $F_{(bs,k[i])}$ in the condition Eq.(29) is shown as follows.

$$\frac{\mathrm{d}d_{(bs,j[i])}}{\mathrm{d}F_{(bs,k[i])}} = -\frac{\mathrm{d}d_{(bs,j[i])}}{\mathrm{d}F_{(bs,j[i])}} < 0 \tag{31}$$

Consequently, in the condition Eq.(29), $d_{(bs,j[i])}$ is a monotonically decreasing function of $F_{(bs,k[i])}$. Because Eq.(21) can be applied to M-route for download traffic, $m_{bs[i]}$ which is the cost of M-route to terminal i is the sum of $d_{(bs,k[i])}$, which is a monotonically increasing function of $F_{(bs,k[i])}$, and $d_{(bs,j[i])}$, which is a monotonically decreasing function of $F_{(bs,k[i])}$, and the 11-downlink cost of the others, which is constant for $F_{(bs,k[i])}$. Therefore, $m_{bs[i]}$ is a multioptimization function of $F_{(bs,k[i])}$, and it has some local minimums for $F_{(bs,k[i])}$ (see Fig. 6(b)).

Here, argue the dependence of $m_{bs[i]}$ on $F_{(bs,k[i])}$. it is shown as follows using Eq.(30) and (31).

$$\frac{\mathrm{d}m_{bs[i]}}{\mathrm{d}F_{(bs,k[i])}} = \frac{\mathrm{d}d_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}} - \frac{\mathrm{d}d_{(bs,j[i])}}{\mathrm{d}F_{(bs,j[i])}} \tag{32}$$

Furthermore, Eq.(32) is transformed into difference equation, and is applied finite difference approximation based on Eq.(29), then the condition that the M-route cost for download traffic decreases is shown as follows.

$$\frac{\mathrm{d}m_{bs[i]}(n)}{\mathrm{d}F_{(bs,k[i])}} \approx \overline{T}_{(bs,k[i])}(n+1) - \overline{T}_{(bs,j[i])}(n+1) > 0 \tag{33}$$

Where, $m_{bs[i]}(n)$ is the cost of M-route to terminal i from base station in packet distribution of n time and $\overline{T}_{(bs,y[x])}(n+1)$ is average delay of interface y in packet distribution of $n+1$ time, and the interface y provides 11-downlink to terminal x.

Furthermore, based on Eqs. (29) and (30), the dependence of $T_{(bs,j[i])}$ on $F_{(bs,k[i])}$ is shown as follows.

$$\frac{\mathrm{d}\overline{T}_{(bs,j[i])}}{\mathrm{d}F_{(bs,k[i])}} < 0 \tag{34}$$

That is, $\overline{T}_{(bs,j[i])}$ is a monotonically decreasing function of $F_{(bs,k[i])}$. Therefore, a cost of each link in M-route should be considered a monotonically increasing function of the packet arrival rate, and the cost of M-route is the sum of each link cost, is a multioptimization function of $F_{(bs,k[i])}$ (see Fig.6(b)). That is, $m_{bs[i]}$ has some local minimums for $F_{(bs,k[i])}$ and the packet distribution meeting Eq.(33) may not bring $m_{bs[i]}$ to the optimal solution.

On the other hand, $\overline{T}_{(bs,k[i])}$ and $\overline{T}_{(bs,j[i])}$ is respectively a monotonically increasing/decreasing function for $F_{(bs,k[i])}$, and then, in $0 \leq F_{(bs,k[i])} \leq F_{bs[i]}$, the number of solutions which makes $\overline{T}_{(bs,k[i])}$ equal to $\overline{T}_{(bs,j[i])}$ is 1 in the maximum (see Fig. 6(c)). Consequently, the packet distribution which meets Eqs.(29) and (33) is repeated, and finally it reaches $\overline{T}_{(bs,k[i])}(n+1) - \overline{T}_{(bs,j[i])}(n+1) = 0$, then the M-route cost $m_{bs[i]}$ reaches its optimal solution. Furthermore, the search for the optimal solution of M-route cost $m_{bs[i]}$ has the additional effectiveness which decreases the arrival of out-of-order packets because of the equalization the delay of source 11-link and destination 11-link.

3. Characteristics of IEEE802.16 link for packet distribution

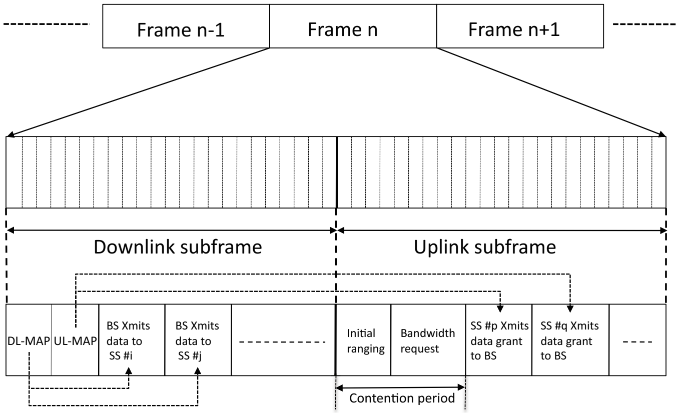


Fig. 7. IEEE802.16 MAC frame.

The performance of IEEE802.16 is actively analyzed. (Nakaya & Hossain, 2006) investigates the delay analysis based on queueing theory, but it does not consider MAC of IEEE802.16. (Cho et al., 2005; Lin et al., 2007; Iyengar et al., 2005; He et al., 2007; Ni et al., 2007) investigate the performance analysis based on MAC of IEEE802.16. Cho et al. (2005) analyzes the utilization and throughput and (Lin et al., 2007) analyzes the utilization for BW request based on polling. These analyses do not investigate the delay. On the other hand, (Iyengar et al., 2005; He et al., 2007; Ni et al., 2007) analyze the delay, but does not consider waiting time in queue. In this section, in regard with IEEE802.16 link (16-link), considering the waiting time in queue and MAC of IEEE802.16, the dependence of average delay on traffic is analyzed in accordance with its four QoS classes. Furthermore, based on the analyzed dependence, the characteristics of 16-link for packet distribution is shown.

Figure. 7 shows 16-frame in TDD. The frame consists of DL-subframe and UL-subframe. Each subframe consists of time slots. Base station (BS) sends DL-MAP and UL-MAP in DL-subframe, and all terminals listen to the DL-subframe, and know that they should listen to slots in DL-subframe, and know that they should use slots in UL-frame to transmit data. In such communications between BS and terminals, IEEE802.16(IEEE std. 802.16-2004, 2004; IEEE std. 802.16e-2005, 2005) supports four class for QoS, which are UGS, rtPS, nrtPS, BE. In UGS class, BS assigns fixed-size periodic data grants to both of uplink and downlink in terminals. In rtPS class and nrtPS class, BS assigns data grants to downlink, and polls to terminals in accordance with the reserved capacity for uplink in each terminal, and in nrtPS class, terminals are additionally allowed to use contention requests for uplink bandwidth (BW). In BE class, terminals are allowed to use contention requests only for both of uplink and downlink, and BS does not poll to terminals.

On the analysis, the assumptions are as follows.

- 16-frame length is constant.
- The multiplexing is TDD.
- The DL-subframe and DL-subframe length in frame is the ratio of 1:1.
- The modulation for each link is unchanged after the communication is arranged
- A time is normalized by slot.

3.1 16-link in UGS

In UGS class, BS assigns fixed-size periodic data grants to both of uplink and downlink in terminals. The fixed-sized periodic data grants is slots of which map is in DL-MAP or UL-MAP. The data arrival process at slot can be approximated to poisson process(Bertsetkas & Gallager, 1992) (Note. data arrival at link means transmission data occurrence in link). Based on the above, argue the average time that a packet waits in queue of downlink, which is $W_{dl.UGS}$. $W_{dl.UGS}$ consists of the follows.

- The average residual time $R_{dl.USG}$. When a new packet arrives at 16-downlink, a 16-frame is already being processed. $R_{dl.USG}$ is a remaining average time until the current 16-frame is processed completely.
- The queued packet average processing time for UGS of downlink, $Q_{dl.UGS}$. $\overline{Q}_{dl.UGS}$ is a average time to process the all queued packets in UGS of downlink on a packet arrival.
- The average advance time $A_{dl.USG}$. In 16-frame, $A_{dl.USG}$ is a average time to process the other packets before a packet in USG of downlink is processed.

$R_{dl.USG}$ consists of $R_{ds.UGS}$, which is the average residual time for the packet in USG of downlink, and R_{other}, which is the average residual time for the packet in frame except for UGS of downlink. Let $C_{dl.UGS}$ be the reserved slots in frame for UGS of downlink, $R_{ds.USG}$ is $C_{dl.UGS}/2$. Let $\overline{V}_{dl.UGS}$ and $\overline{V^2}_{dl.UGS}$ be respectively the first and second moment of process time for a packet in frame except for UGS of downlink, R_{other} is $\overline{V^2}_{dl.UGS}/2\overline{V}_{dl.UGS}$. Let L_F be the number of slots in 16-frame, then $R_{dl.UGS}$ is derived as follows.

$$\begin{aligned} R_{dl.USG} &= \frac{C_{dl.USG}}{L_F} R_{ds.UGS} + \left(1 - \frac{C_{dl.USG}}{L_F}\right) \overline{R}_{other} \\ &= \frac{C^2_{dl.UGS}}{2L_F} + (L_F - C_{dl.UGS}) \frac{\overline{V^2}_{dl.UGS}}{2\overline{V}_{dl.UGS} L_F} \end{aligned} \tag{35}$$

Argue $Q_{dl.UGS}$. Based on Little's theorem(Gross & Harris, 1985), the number of queued packets in UGS of downlink, which is $N_{dl.UGS}$, is derived as follows.

$$N_{dl.UGS} = F_{dl.UGS} \cdot W_{dl.UGS} \tag{36}$$

Where $F_{dl.UGS}$ is a packet arrival rate at UGS of downlink, which is average number of arrival packets within a slot in UGS of downlink, $W_{dl.UGS}$ is the average time that a packet waits in queue in UGS of downlink. Let m be a data grants period which is expressed by the number of frames, and $Q_{dl.UGS}$ is derived as follows.

$$Q_{dl.UGS} = F_{dl.UGS} \cdot W_{dl.UGS} \cdot m \cdot L_F \tag{37}$$

$A_{dl.USG}$ is equal to the residual time of DL-subframe, and is derived as follows.

$$A_{dl.UGS} = \frac{C^2_{dl.UGS}}{2L_{dl}} + (L_{dl} - C_{dl.UGS})\frac{\overline{V^2}_{ds.UGS}}{2\overline{V}_{ds.UGS}L_{dl}} \tag{38}$$

L_{dl} is the number of slots in DL-subframe, $\overline{V}_{ds.UGS}$ and $\overline{V^2}_{ds.UGS}$ are respectively the first and second moment of process time of a packet in DL-subframe except for UGS. Accordingly, $W_{dl.UGS}$ is expressed as follows.

$$\begin{aligned} W_{dl.UGS} &= R_{dl.USG} + F_{dl.UGS} \cdot W_{dl.UGS} \cdot m \cdot L_F + A_{dl.UGS} \\ W_{dl.UGS} &= \frac{R_{dl.UGS} + A_{dl.UGS}}{1 - mF_{dl.UGS}L_F} \end{aligned} \tag{39}$$

Based on Eq. (39), the average delay in UGS of downlink, which is $T_{dl.UGS}$, is derived as follows.

$$T_{dl.UGS} = W_{dl.UGS} + C_{dl.UGS} \tag{40}$$

Assuming the modulation for each link to be unchanged, $C_{dl.UGS}$, $\overline{V}_{dl.UGS}$, $\overline{V^2}_{dl.UGS}$, $\overline{V}_{ds.UGS}$, and $\overline{V^2}_{ds.UGS}$ are constant even if $F_{dl.UGS}$ changes, and they are independent on $F_{dl.UGS}$. That is, $R_{dl.UGS}$ and $A_{dl.UGS}$ are independent on $F_{dl.UGS}$. Therefore, using Eq. (40), the first and second derivative of $T_{dl.UGS}$ at $F_{dl.UGS}$ are derived respectively as follows.

$$\frac{\mathrm{d}T_{dl.UGS}}{\mathrm{d}F_{dl.UGS}} > 0 \qquad \frac{\mathrm{d}^2T_{dl.UGS}}{\mathrm{d}F^2_{dl.UGS}} > 0 \tag{41}$$

Consequently, $T_{dl.UGS}$ is a convex monotonically increasing function of $F_{dl.UGS}$.

Argue $W_{ul.UGS}$, which is the average time that a packet waits in queue of uplink. Similar to $W_{dl.UGS}$, $W_{ul.UGS}$ consists of $R_{ul.UGS}$, which is the average residual time for frame on a packet arrival at USG of uplink, $Q_{ul.UGS}$, which is the queued packet processing time for UGS of uplink, and $A_{ul.UGS}$ which is the average advance time for UGS of uplink. $R_{ul.UGS}$ is common to $R_{dl.UGS}$, and $Q_{ul.UGS}$ is $F_{ul.UGS}W_{ul.UGS}mL_F$ based on Little's theorem. $A_{ul.UGS}$ is the sum of L_{dl} and the residual time for UL-subframe because UL-subframe is arranged to following DL-subframe. Let $T_{ul.UGS}$ and $\overline{C}_{ul.UGS}$ be respectively the average delay in USG of uplink and the number of reserved slots for UGS of uplink, $W_{ul.UGS}$ and $T_{dl.UGS}$ are respectively dervied as follows

$$\begin{aligned} W_{ul.UGS} &= \frac{R_{ul.UGS} + A_{ul.UGS}}{1 - mF_{ul.UGS}L_F} \\ T_{ul.UGS} &= W_{ul.UGS} + C_{ul.UGS} \end{aligned} \tag{42}$$

Similar to downlink, $R_{ul.UGS}$ and $A_{ul.UGS}$ are independent on $F_{ul.UGS}$. Accordingly, $T_{ul.UGS}$ is a convex monotonically increasing function of $F_{ul.UGS}$.

3.2 16-downlink in rtPS and nrtPS

In rtPS, BS periodically assigns data grants to downlink of terminals based on the reserved capacity for the link. Similar to UGS, delay of 16-downlink in rtPS, which is $T_{dl.rtPS}$, is derived as follows.

$$\begin{aligned}
R_{dl.rtPS} &= \frac{\overline{X^2}_{dl.rtPS}}{2L_F} + (L_F - \overline{X}_{dl.rtPS})\frac{\overline{V^2}_{dl.rtPS}}{2\overline{V}_{dl.rtPS}L_F} \\
A_{dl.rtPS} &= \frac{\overline{X^2}_{dl.rtPS}}{2L_{dl}} + (L_{dl} - \overline{X}_{dl.rtPS})\frac{\overline{V^2}_{ds.rtPS}}{2\overline{V}_{ds.rtPS}L_{dl}} \\
W_{dl.rtPS} &= \frac{R_{dl.rtPS} + A_{dl.rtPS}}{1 - mF_{dl.rtPS}L_F} \\
T_{dl.rtPS} &= W_{dl.rtPS} + \overline{X}_{dl.rtPS} \\
\overline{X}_{dl.rtPS} &+ \overline{V}_{dl.rtPS} = L_F \\
\overline{X}_{dl.rtPS} &+ \overline{V}_{ds.rtPS} = L_{dl}
\end{aligned} \tag{43}$$

$\overline{X}_{dl.rtPS}$ and $\overline{X^2}_{dl.rtPS}$ are respectively the first and second moment of the number of granted slots, which is a process time of a packet, for rtPS of downlink, $\overline{V}_{dl.rtPS}$ and $\overline{V^2}_{dl.rtPS}$ be respectively the first and second moment of process time of a packet in frame except for rtPS of downlink, $\overline{V}_{ds.rtPS}$ and $\overline{V^2}_{ds.rtPS}$ be respectively the first and second moment of process time of a packet in DL-subframe except for rtPS, $F_{dl.rtPS}$ is a rtPS packet arrival rate at 16-downlink, and $W_{dl.rtPS}$ is the average time that a packet waits in queue in rtPS of downlink,

Argue the dependence of $\overline{X}_{rtPS}$ and $\overline{X^2}_{dl.rtPS}$ on $F_{dl.rtPS}$. Assuming the modulation for each link to be unchanged, $\overline{X}_{rtPS}$ increases in the linear for the increase in $F_{dl.rtPS}$. Therefore, the dependence of $\overline{X}_{rtPS}$ and $\overline{X^2}_{dl.rtPS}$ on $F_{dl.rtPS}$ are respectively expressed as follows.

$$\begin{aligned}
\frac{\mathrm{d}\overline{X}_{dl.rtPS}}{\mathrm{d}F_{dl.rtPS}} > 0 \qquad \frac{\mathrm{d}^2\overline{X}_{dl.rtPS}}{\mathrm{d}F^2_{dl.rtPS}} = 0 \\
\frac{\mathrm{d}\overline{X^2}_{dl.rtPS}}{\mathrm{d}F_{dl.rtPS}} > 0 \qquad \frac{\mathrm{d}^2\overline{X^2}_{dl.rtPS}}{\mathrm{d}F^2_{dl.rtPS}} = 0
\end{aligned} \tag{44}$$

Based on Eq. (43) and (44), the dependence of $T_{dl.rtPS}$ on $F_{dl.rtPS}$ is derived as follows.

$$\frac{\mathrm{d}T_{dl.rtPS}}{\mathrm{d}F_{dl.rtPS}} > 0 \qquad \frac{\mathrm{d}^2T_{dl.rtPS}}{\mathrm{d}F^2_{dl.rtPS}} > 0 \tag{45}$$

The difference of nrtPS form rtPS is the length of data grant periods, and the data grants period in nrtPS is longer than that in rtPS. Then the depenadence of delay $T_{dl.nrPS}$ on $F_{dl.nrtPS}$, which is nrtPS packet arrival rate at uplink, is the same as that in rtPS. Consequently, $T_{dl.rtPS}$ and $T_{dl.nrtPS}$ are a convex monotonically increasing function of the each packet arrival rate.

3.3 16-uplink in rtPS

In rtPS, BS periodically polls to terminals in accordance with the reserved capacity for uplink, and terminals reply by sending BW requests with allocated space (i.e., contention free). In next frame, BS assigns data grants which is mapped by UL-MAP to terminals, and terminals use data grant to transmit data. The difference of rtPS of uplink from that of downlink is that two frames is necessary to transmit a packet. Let $R_{ul.rtPS}$, $A_{ul.rtPS}$, m, and $F_{ul.rtPS}$, $\overline{X}_{ul.rtPS}$ be respectively the average residual time for rtPS packet of uplink, the average advance time for rtPS packet of uplink, the polling period in rtPS, the packet arrival rate at rtPS of uplink, and the average process time for packet in rtPS of uplink, $W_{ul.rtPS}$, which is the queued packet processing time for UGS of uplink, and $T_{ul.rtPS}$, which is the average delay in rtPS of uplink, are respectively expressed as follows.

$$\begin{aligned} W_{ul.rtPS} &= \frac{R_{ul.rtPS} + A_{ul.rtPS}}{1 - 2mF_{ul.rtPS}L_F} \\ T_{ul.rtPS} &= W_{dl.rtPS} + \overline{X}_{ul.rtPS} \end{aligned} \tag{46}$$

$R_{ul.rtPS}$ is common to $R_{dl.rtPS}$, and $A_{ul.rtPS}$ is the sum of $A_{dl.rtPS}$ and L_F because the rtPS of uplink is necessary to additional a frame to poll to terminal and to request BW to BS with contention free. Therefore, $R_{dl.rtPS}$ and $A_{dl.rtPS}$ are independence on $F_{ul.rtPS}$, and then $T_{ul.rtPS}$ is a convex monotonically increasing function of $F_{ul.rtPS}$ the same as rtPS of downlink.

3.4 16-uplink in nrtPS and 16-link in BE

In 16-uplink of nrtPS and 16-link of BE, also the arrival packets are enqueued and wait to be processed with FCFS. Let the waiting time be W_{bw} (argue later in detail). The packet is dequeued with FCFS, and then, is processed. The packet processing in nrtPS is based on the polling from BS the same as uplink of rtPS. Furthermore, uplink of nrtPS is additionally allowed to use contention BW request. In BE, the link is allowed to use contention BW request only. In such contention mode, terminals send BW request during the contention period in UL-subframe. Depending on the number of contention BW request, the collision of BW request occurs. In contention BW request, each terminal resolves and avoids the collision as follows.

- Each terminal waits the random number of slots before sending BW request in the contention period. The number of waiting slots, which is back-off counter, is generated based on exponential binary backoff mechanism.
- The backoff counter is decreased during the contention period.
- When the counter is zero, terminal sends BW request in the contention period.
- The terminal sending BW request waits data grants in DL/UL-map from BS.
- When the terminal does not receive data grants from BS in duration of the timer, terminal increases the contention window size, and generates the backoff counter based on exponential binary backoff mechanism, and then waits the opportunity sending BW request when the counter is zero. That is the retransmission process.

The contention BW request is analyzed based on the following model.

- The packet processing time consists of BW request opportunity waiting period, BW request attempt period, and packet transmission period.
- A BW request opportunity waiting period is the number of slots to be spent until the back-off counter becomes zero.
- A BW request attempt period is the number of slots to be spent by BW request transmission. In BW request attempt period, BW request transmission succeeds or collides. The collision causes the timeout in receipt of data grant, and spends the number of slots corresponding to the timeout. The success spends the number of slots to be spent from BW request accepted by BS to complete transmission of a packet in terminal.

- In each terminal, Let τ_{bw} be the BW request attempt rate (req/slot) in the contention period of UL-subframe, and then the probability q_{bw} that BW request is transmitted successfully is $(1 - \tau_{bw})^{n-1}$, where n is the number of terminals transmitting BW request .
- In each terminal, the packet arrival process (i.e., upload traffic) and packet request process (i.e., download traffic) is poisson process(Bertsetkas & Gallager, 1992). Let F_{bw} be an packet arrival/request rate (packets/slot), which need the contention BW request.
- The contention period ratio, which is the ratio of the number of slots in the contention period in a frame, is constant. Let U_c be the contention period ratio.
- The process of the BW request that BS receives is assumed to FCFS, and the allocating data grants rate (slot/packet) in DL-subframe or UL-subframe for BW request in BS is S_{dg}, and is constant.

The contention BW request process is the same as the model described in 2.1.1 except for t_b in Eq.(4), t_s and t_c in Eq.(6). t_b is 1 because the contention BW request process decrements the backoff counter without carrier sensing. t_c is the number of slots to be spent by timeout of data grant receipt from BS, and is a constant. t_s is the number of slots to be spent from the success transmitting of BW request to the complete transmission of packet, and then it depends on F_{bw}. t_s is divided into t_{ss}, which is the air time of BW request from terminal to BS, and t_{bs}, which is the time from the receipt of BW request in BS to the complete transmission of packet in terminal, and t_{ss} is a constant.
Here, argue the dependence of t_{bs} on F_{bw}. In S_{bw} which is the average time from first transmission attempt of contention BW request to successful transmission of that, the average number of arrival/request packets for contention BW request is $F_{bw}S_{bw}$, and, in S_{bw}, the average number of BW request transmission attempts is $(1 - q_{bw})/q_{bw} + 1$. Therefore, τ_{bw} is expressed as follows.

$$\tau_{bw} = \frac{F_{bw}S_{bw}}{U_c S_{bw} q_{bw}} = \frac{F_{bw}}{U_c q_{bw}} \tag{47}$$

And, based on Eqs.(3), (4) and (5), S_{bw} is shown as follows.

$$\begin{aligned} S_{bw} &= B_{bw} + A_{bw} \\ B_{bw} &= \tfrac{\eta CW_{min}-1}{2q} \\ A_{bw} &= \tfrac{1-q}{q} t_c + t_{ss} \end{aligned} \tag{48}$$

Furthermore, let F_{bw_bs} be the arrival rate of BW request at BS, F_{bw_bs} is shown as follows.

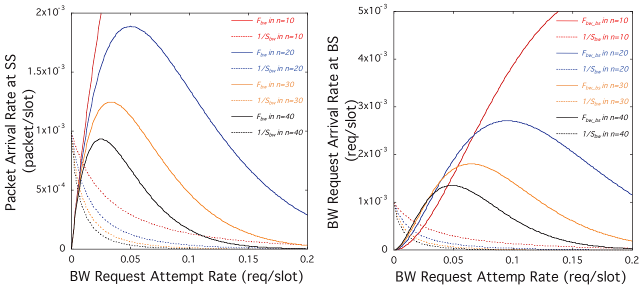
$$F_{bw\_bs} = q_{bw} n F_{bw} \tag{49}$$

Based on Eqs.(47), (48) and (49), on condition of $F_{bw} < 1/S_{bw}$, the dependence of F_{bw}, F_{bw_bs} and S_{bw} on τ_{bw} is respectively shown as follows.

$$\frac{\mathrm{d}F_{bw}}{\mathrm{d}\tau_{bw}} > 0 \qquad \frac{\mathrm{d}F_{bw\_bs}}{\mathrm{d}\tau_{bw}} > 0 \qquad \frac{\mathrm{d}S_{bw}}{\mathrm{d}\tau_{bw}} > 0 \tag{50}$$

Figure 8(a) and 8(b) respectively illustrates the dependence of F_{bw} and F_{bw_bs} on τ_{bw} by using Eqs. (47), (49), and each also shows the same characteristics. Therefore, on condition of $F_{bw} < 1/S_{bw}$, the dependence of F_{bw_bs} and S_{bw} on F_{bw} is respectively shown, by using Eq.(50), as follows.

$$\frac{\mathrm{d}F_{bw\_bs}}{\mathrm{d}F_{bw}} > 0 \qquad \frac{\mathrm{d}S_{bw\_bs}}{\mathrm{d}F_{bw}} > 0 \tag{51}$$



(a) Dependence of F_{bw} on τ_{bw}. (b) Dependence of F_{bw_bs} on τ_{bw}.

Fig. 8. Dependence of each element on τ_{bw}.

Argue W_{bw_bs} which is the waiting time in queue of BS for data grant. The process of the received BW requests in BS is assumed to be FCFS, and conceptually it can be view as queueing system of which the packet arrival rate is F_{bs_bs} and the packet service rate is S_{dg}. Therefore, W_{bw_bs} is shown, based on Eq.(14), as follows

$$W_{bw\_bs} = \frac{F_{bw\_bs}\overline{S_{dg}^2}}{2(1 - F_{bw\_bs}S_{dg})} \tag{52}$$

S_{dg} is constant for F_{bw_bs}, and then, on condition of $F_{bw\_bs} < 1/S_{dg}$, the dependence of W_{bw_bs} on F_{bw_bs} is shown as follows.

$$\frac{\mathrm{d}W_{bw\_bs}}{\mathrm{d}F_{bw\_bs}} > 0 \tag{53}$$

t_{bs} is the sum of W_{bw_bs} and S_{dg}, and then, based on Eqs.(51) and (56), the dependence of t_{bs} on F_{bw} on condition of $F_{bw} < 1/S_{bw}$ and $F_{bw\_bs} < 1/S_{dg}$, that is, within the link capacity, is shown as follows.

$$\frac{\mathrm{d}t_{bs}}{\mathrm{d}F_{bw}} > 0 \tag{54}$$

Argue W_{bw} which is the packet waiting time in queue of terminal. According to the exponential binary backoff model described in 2.1.1, and applying $t_b = 1$ and the constance of t_c for τ_{bw} to Eq.(15), W_{bw} is derived as follows.

$$W_{bw} = \frac{F_{bw}\overline{S_{bw}^2}}{2(1 - F_{bw}S_{bw})} \tag{55}$$

According to Eqs.(16), (50), on condition of $F_{bw} < 1/S_{bw}$, that is, within link capacity, the dependence of W_{bw} on F_{bw} is derived as follows.

$$\frac{\mathrm{d}W_{bw}}{\mathrm{d}F_{bw}} > 0 \tag{56}$$

Finally, T_{bw}, which is the average delay for contention BW request, is the sum of W_{bw}, S_{bw}, t_{ss} and t_{bs}, the dependence of T_{bw} on F_{bw} is derived, based on Eqs.(51), (54) and (56), as follows.

$$\frac{\mathrm{d}T_{bw}}{\mathrm{d}F_{bw}} = \frac{\mathrm{d}}{\mathrm{d}F_{bw}}(W_{bw} + S_{bw} + t_{ss} + t_{bs}) > 0 \tag{57}$$

Therefore, T_{bw} is monotonically increasing function of F_{bw}.

3.5 Packet distribution for 16-link

The average delay on 16-link, except uplink in nrtPS and link in BE, is a convex monotonically increasing function of packet arrival rate, therefore, its characteristics on packet distribution corresponds to that of 11-downlink. On the other hand, 16-uplink in nrtPS and 16-link in BE are a monotonically increasing function of packet arrival rate, therefore, their characteristics on packet distribution corresponds to that of 11-uplink.

4. IP packet distribution for M-route compositing IEEE802.11/16 links

Based on the analyzed characteristics of 11/16-link for packet distribution, the characteristics of the access route compositing multiple 11-links or 16-links is the same. Therefore, the characteristics of M-route compositing 11links and 16-links for the packet distribution corresponds to that of the access route compositing multiple 11-links or 16-links.

According to the above, IP packet distribution method for the M-route compositing 11-links and 16-links be described.

The characteristics of M-route compositing 11/16-link for the packet distribution corresponds to that of the access route compositing multiple 11-links or 16-links because that of 11-link and 16-link are the same.

4.1 Restriction condition

According to Eqs.(26) and (33), the optimal solution of the M-route cost can be searched by the repeating packet distribution that the average delay of distribution source link becomes larger than that of distribution destination link, and that the average delay of both source link and destination link become equal finally. Additionally, the packet distribution for download traffic needs to meet the condition Eq.(29) when the source link on the packet distribution is an 11-link.

Here, argue the condition Eq.(29). Transforming Eq.(29) into finite difference approximation, it is shown as follows.

$$\frac{\mathrm{d}\overline{T}_{(bs,k[i])}}{\mathrm{d}F_{(bs,k[i])}} \approx \frac{\Delta\overline{T}_{(bs,k[i])}}{\Delta F_{(bs,k[i])}} > 0 \tag{58}$$

Because 11-link $k[i]$ is a source link on packet distribution, $\Delta F_{(bs,k[i])} < 0$. Therefore, to meet Eq.(58), $\Delta\overline{T}_{(bs,k[i])} < 0$. In other words, it is that the average delay of source 11-interface on packet distribution decreases. The increase in average delay of source 11-interface k does not meet the condition and it occurs in the following unsteady state.

- The packet arrival rate at other links provided by 11-interface k increases.
- The number of links provided by 11-interface k increases.

The first item in the above list means the increase in contention with other terminals, thus it also causes the increase in average delay of source link when source link is 11-uplink or 16-uplink in nrtPS or 16-link in BE. The second item means the increase in a number of terminals, thus it causes the increase in average delay of source link because of the same reason as the first item. Then it also occurs when source link on packet distribution is 11-uplink or 16-uplink in nrtPS or 16-link in BE. In above cases, M-route cost also loses the monotonically increasing characteristics for packet arrival rate. Therefore, in consideration of the unsteady state that traffic fluctuates, the restriction condition which is the decrease in the average delay of source link on packet distribution is a necessary condition to bring the M-route cost to the optimal solution.

4.2 Search for optimal solution of M-route cost with packet distribution

Argue the search for optimal solution of M-route cost with Packet Distribution in unsteady state by the following packet distribution.

- $M_{(x,y)}$ is a M-route from x to y. On x and y, one is a base station and the other is a terminal.
- Packets transmitted to y at x are distributed.
- K denotes a source interface on the packet distribution. J denotes a destination interface on the packet distribution.
- K is either an 11-interface or 16-interface and J is also either an 11-interface or 16-interface.
- $(x, Z[y])$ denotes a certain link to y in x, which link is provided by a certain interface Z.
- $F_{(x,Z[y])}$ denotes a packet arrival rate at $(x, Z[y])$
- $T^*(p_{(x,Z[y])})$ denotes interface average delay $\overline{T}(p_{(x,Z[y])})$ if Z is 11-interface, and denotes link average delay $T(p_{(x,Z[y])})$ if Z is 16-interface.

Based on subsection 4.1, the search for optimal solution of M-route cost in unsteady state is the search for the packet distribution meeting the following conditions.

$$T^*_{(x,K[y])}(n) - T^*_{(x,J[y])}(n) > 0 \qquad \Delta T^*_{(x,K[y])}(n) < 0 \tag{59}$$

where $\Delta T^*_{(x,K[y])}(n)$ denotes the difference between $T^*_{(x,K[y])}(n)$ and $T^*_{(x,K[y])}(n-1)$. According to Eq. (59), the proposed packet distribution method implements the search for the optimal solution in IP layer using the measured average delay in MAC layer as the following iteration.

Step1: In the initial period, packets are distributed equally to each link in M-route with a round robin manner.

Step2: At end of the initial period, $T^*_{(x,Z[y])}(0)$ of each link in M-route is derived, and $(x, Max[y])(0)$ which has maximum average delay in the initial (0-th) period, and $(x, Min[y])(0)$ which has minimum average delay in the initial (0-th) period, is respectively selected in $M_{(x,y)}$. On the packet distribution, $(x, Max[y])(0)$ and $(x, Min[y])(0)$ is respectively assigned to the source link $(x, K[y])(1)$ in the next (1-th) period and the destination link $(x, J[y])(1)$ in that period. $\Delta F_{(x,K[y])}(1)$, which is the amount of packet distribution from $(x, K[y])(1)$ to $(x, J[y])(1)$ in the next (1-th) period, is derived as follows. where $r_{(x,y)}$ denotes the packet distribution rate of $M_{(x,y)}$, and $r0$ denotes the initial packet distribution rate.

$$\begin{aligned} &\Delta F_{(x,K[y])}(1) = r_{(x,y)}(1) \cdot F_{(x,K[y])}(0) \\ &r_{(x,y)}(1) = r0 \end{aligned} \tag{60}$$

Step3: According to $\Delta F_{(x,K[y])}(1)$, the packet distribution in the 1-th period is carried out.

Step4: At end of n-th period ($n \geq 1$), $T^*_{(x,Z[y])}(n)$ of each link in $M_{(x,y)}$ is derived. The delay of each packet is a period when the packet arrives at IP layer, and is enqueued in queue of an interface, and is dequeued by an interface, and is sent and interface receives its ACK based on the media access control. Therefore, it can be measured within packet distributing side x. Based on the relation of $T^*_{(x,K[y])}(n)$ and $T^*_{(x,J[y])}(n)$, $\Delta F_{(x,K[y])}(n+1)$ is derived as follows.

- In $T^*_{(x,K[y])}(n) > T^*_{(x,J[y])}(n)$ and in $\Delta T^*_{(x,K[y])}(n) < 0$, Eq. (59) is met. Therefore, $\Delta F_{(x,K[y])}(n+1)$ is allocated the same as $\Delta F_{(x,K[y])}(n)$, and it is shown as follows.

$$\begin{aligned} \Delta F_{(x,K[y])}(n+1) &= r_{(x,y)}(n+1) \cdot \Delta F_{(x,K[y])}(n) \\ r_{(x,y)}(n+1) &= r_{(x,y)}(n) \end{aligned} \tag{61}$$

- In $T^*_{(x,K[y])}(n) < T^*_{(x,J[y])}(n)$ and in $\Delta T^*_{(x,K[y])}(n) < 0$, $M_{(x,y)}$ cost goes beyond the optimal solution and ascents the gradient. Because it is caused by the excessive packet distribution from source link to destination link, $\Delta F_{(x,K[y])}(n+1)$ is allocated smaller than $\Delta F_{(x,K[y])}(n)$ as follows. where α is the decrement rate ($0 < \alpha < 1$).

$$\begin{aligned} \Delta F_{(x,K[y])}(n+1) &= r_{(x,y)}(n+1) \cdot \Delta F_{(x,K[y])}(n) \\ r_{(x,y)}(n+1) &= \alpha \cdot r_{(x,y)}(n) \end{aligned} \tag{62}$$

- In $\Delta T^*_{(x,K[y])}(n) > 0$, the traffic among the source link increases as shown in subsection 4.1. Because $\Delta F_{(x,K[y])}(n)$ is underestimated, and because the monotonically increasing characteristics of the source link cost for the packet distribution is regained, $\Delta F_{(x,K[y])}(n+1)$ is allocated larger than $\Delta F_{(x,K[y])}(n)$ as follows. where β is the increment rate ($\beta > 1$).

$$\begin{aligned} \Delta F_{(x,K[y])}(n+1) &= r_{(x,y)}(n+1) \cdot \Delta F_{(x,K[y])}(n) \\ r_{(x,y)}(n+1) &= \beta \cdot r_{(x,y)}(n) \end{aligned} \tag{63}$$

Step5: $(x, Max[y])(n)$ and $(x, Min[y])(n)$ are respectively selected in $M_{(x,y)}$, and are respectively assigned to $(x, K[y])(n+1)$ and $(x, J[y])(n+1)$. According to $(x, K[y])(n+1)$, $(x, J[y])(n+1)$, and $\Delta F_{(x,K[y])}(n+1)$, the $(n+1)$-th packet distribution is carried out, then return to Step4.
In each M-route of both a base station and terminals, the above iteration gradually updates the amount of packet distribution, and brings M-route cost to the optimal solution, reducing the out-of-order packets occurred by distributing packets to multiple links.

5. Performance evaluation

In this section, the simulation evaluation of the packet distribution method for M-route compositing 11/16-links is shown.

5.1 Simulation scenario

For the simulation evaluation, OPNET 12.0A PL3 was used, and the network configuration was as follows (see Fig. 9):

- Base station is equipped with an 16-interface and 4×11a/b-interfaces. 16-interface and 11a/b-interface respectively connects to 16-antenna and 11/ab-antenna.
- The number of terminals is 100, and each terminal is equipped with 16-interface and 11a/b-interface.

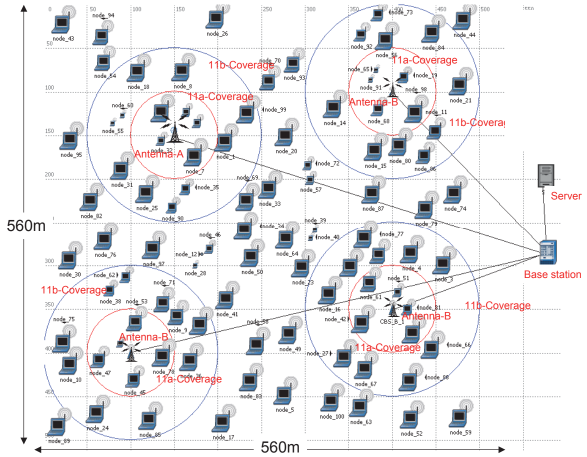


Fig. 9. Example of access network topology.

- An antenna-A which equips with 16- and 11a/b-antena, three antenna-B which equip with 11a/b-antena, and 100 terminals without mobility are randomly deployed in 560m × 560m space with a 1/10 scale of 16-coverage with 1000m radius.
- A FTP server and a Video Conference (VC) server, which are outside the wireless access network, are connected to the base station by a wired network.

In the above access network, M-route between each terminal and a base station combines available links as follows.

- The M-route between a base station and a terminal in 11a-coverage (area-A) combines 11a/b-link and 16-link.
- The M-route between a base station and a terminal in 11b-coverage and outside 11a-coverage (area-B)combines 11b-link and 16-link.
- The M-route between a base station and a terminal outside 11b-coverage (area-C) uses only 16-link.

The performance of 11a/b-wireless system and 16-wireless system shown in Table 1 is applied, and each the capacity reservation of 16-link is shown in Table 2. Assuming the evaluation environment to be a suburban area in line of sight, the 11a/b-radio propagation model is a two-ray model and Ricean fading with Ricean factor 6dB(Takada, 2004), and the 16 radio propagation model is a Erceg (TerrainA).
According to (3GPP2, 2006), the VC traffic on UDP is generated at each terminal as follows:

- The average video rate in the incoming and the outgoing is 32 Kbps.
- The distribution in video rate is a truncated pareto distribution with maximum 8Kbits
- The frame rate in the incoming and outgoing is 10fps. A frame corresponds to a data packet in VC.
- As the sequence control of frame, VC waits for the frame with expected sequence number for a period of 100 msec that is equal to frame interval. The frame that arrived on excess of the period is destroyed.

- In 16-link, VC is mapped to rtPS for QoS class.

Furthermore, FTP traffic on TCP is also generated at each terminal as follows:

- In 10 sec period, FTP session which transfers a file of the size of 1K$\sim$400Kbytes starts.
- 50% of the FTP sessions are download session.
- Each FTP session is established between each terminal and a FTP server.
- In 16-link, FTP is mapped to nrtPS for QoS class.

The evaluation items are as followings.

- IP average delay (sec/packet), is the average delay between terminal and servers in an IP packet.
- IP throughput (bps), is the average arrival amount of IP packets at terminals and servers during a unit time.
- FTP response time (sec/file), is the average delay to transfer a file in end-to-end between a terminal and an FTP server.
- FTP throughput (bytes/sec), is the average amount of arrival data packets at terminals and an FTP server during a unit time.
- VC average delay (sec/frame), is the average delay of end-to-end between terminal and a VC server in a data frame.
- VC throughput (bytes/sec), is the average arrival amount of data frames at terminals and a VC server during a unit time.

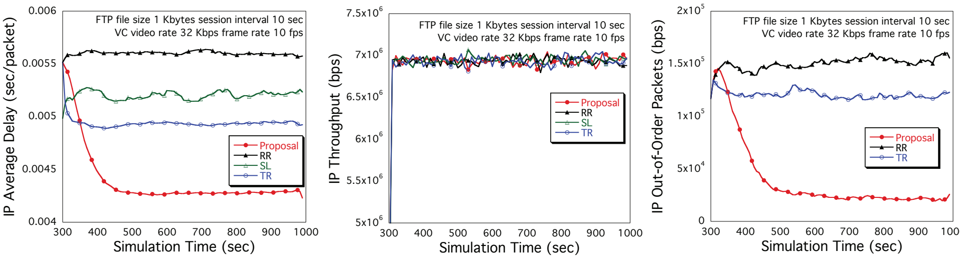
The end-to-end delay is composed of the delay in wireless access network and that in wired communication between the base station and server. The delay in wired communication is common without depending on any packet distribution in wireless access network because the wired communication is out of scope of wireless access network. Therefore, the delay in wired communication can be assumed to be constant to any packet distribution in wireless access network, and the delay in wireless access network depends on packet distribution in wireless access network. In viewpoint of packet distribution, the trend of the end-to-end delay corresponds with that of the delay in wireless access network. Thus the delay in wired communication can be logically ignored. Furthermore, assuming the access speed of a future core network to be Gigabits order(Konishi et al., 2008), the delay in WiFi corresponds to $10^2 \sim 10^3$ order of that in wired core network because the bandwidth of WiFi is Mbps. Then, the delay between the base station and server is left out of consideration because it is independent on the performance of the wireless access network. Furthermore, to demonstrate the effectiveness of the proposed method, it is compared with the following methods.

- Single link (SL) uses a link. The terminals in area-A use 11a-link, the terminals in area-B use 11b-link, and the other terminals use 16-link.
- Round robin (RR) uses available links and distributes packets equally to each link.
- Actual transmission rate (TR) uses available links and distributes packets to each link in proportion to the measured transmission rate at each link in every 10 sec.

In the search for minimal solution, $r0$ is 0.1, α is 0.5, β is 1.5, and the update period of packet distribution is 10 sec.
Furthermore, the link combination in IP is transparent to the upper layer. Therefore, the upper layer is provided with the M-route as a single link view.

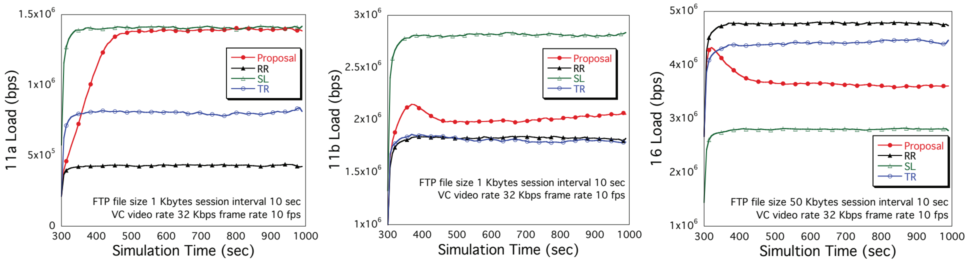
QoS Class	rtPS	nrtPS
Maximum Sustained Transmission Rate	384Kbps	384Kbps
Minimum Reserved Transmission Rate	80Kbps	1Kbps

Table 2. Capacity reservation for 16-link.



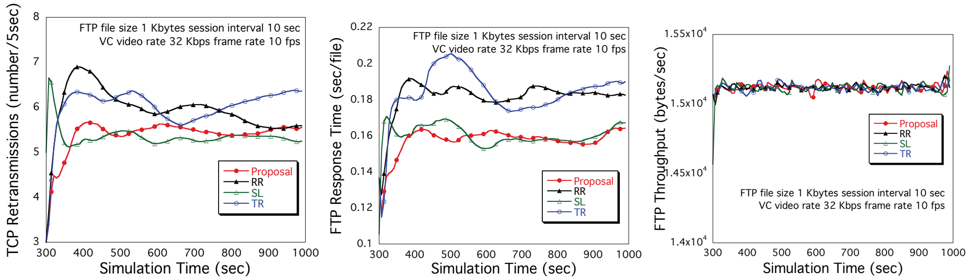
(a) Average delay.
(b) Throughput.
(c) Out-of-order.

Fig. 10. Transition of IP on FTP file size 1K bytes.



(a) 11a load.
(b) 11b load.
(c) 16 load.

Fig. 11. Distributed traffic load to each wireless system on FTP file size 1K bytes.

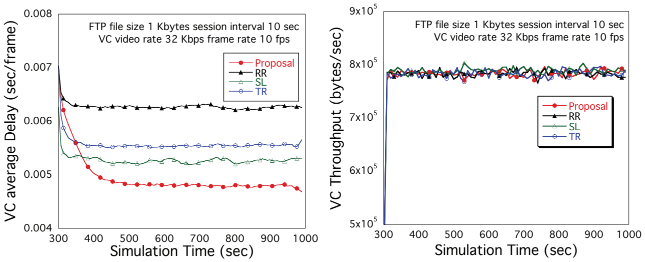


(a) TCP retransmissions.
(b) FTP response time.
(c) FTP throughput.

Fig. 12. Transition of TCP and FTP on FTP file size 1K bytes.

5.2 Transition of delay and throughput in low traffic load

Figures 10(a) and 10(b) show, respectively, the transition of IP average delay and IP throughput, when file size in FTP is 1K bytes. As the packet distribution proceeds, the IP average delay of the proposal decreases rapidly, and becomes much lower than that of the



(a) Average delay.

(b) Throughput.

Fig. 13. Transition of VC on FTP file size 1K bytes.

others. Figures 11(a), 11(b) and 11(c) show, respectively, the transition of distributed load to 11a-wireless system (11a-load), that to 11b-wireless system (11b-load) and that to 16-wireless system (16-load), when file size in FTP is 1K bytes. The decrease in IP average delay of the proposal corresponds to the increase in 11a-load of the proposal (see Fig. 10(a) and Fig. 11(a)). In area-A, 11a accommodates a few terminals because of its narrow coverage, and the proposal distributes almost packets to 11a-link the same as SL, and saves the capacity of 11b and 16 for many terminals outside area-A. RR and TR in the area distributes packets to other link as well, thus RR and TR can not use 11a capacity effectively to save the capacity of 11b and 16. Consequently, RR and TR bring the large load to 16 (see Fig. 11(c)), which of links have low transmission rate (see Tab. 2), and it causes the inferior IP average delay of RR and TR to that of the proposal. In area-B, SL distributes all packets to 11b-link (see Fig. 11(b)), and then the packet collision in 11b occurs frequently. Thus, it causes the inferior IP average delay of SL to that of the proposal. In comparison with SL, the packet distribution of the proposal and TR improve IP performance, but that of RR lowers IP performance.

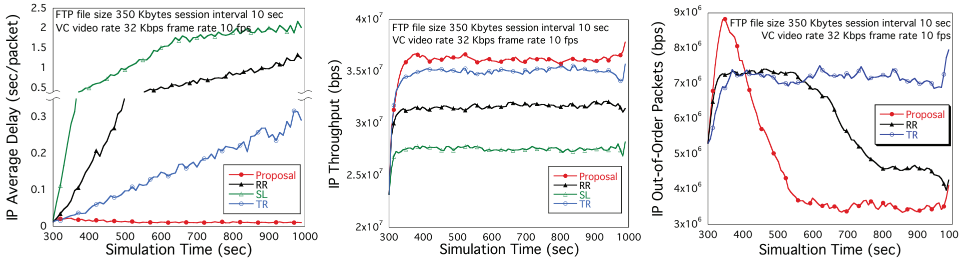
The IP out-of-order packets of the proposal decreases the same as the decrease in its IP average delay, consequently, its out-of-order packets becomes much lower than that of RR and TR (see Fig. 10(c)). Therefore, its packet distribution effects the decrease in IP average delay and the decrease in out-of-order packets. Figures 12(a) shows the number of TCP retransmissions for a period of 5 sec. The TCP retransmissions of the proposal is nearly equal to that of SL and RR, and that of TR is larger than that of the others. The cause of TCP retransmission in SL is packet loss. In area-B, SL distributes all packets to 11b, thus the packet collision occurs frequently in 11b and then it causes the TCP retransmission. The cause of TCP retransmission in the proposal, RR and TR is out-of-oder packets. The number of TCP transmissions in RR is lower than that of TR. RR loads larger mount of packets with 16 than the others (see Fig. 11(c)). Because the 16-link has the low transmission rate, the IP average delay of RR is inferior to that of the others (Fig. 10(a)). Then TCP congestion window size of RR is smaller than that of TR and the proposal, and the amount of distributed packets to multiple links for a period is fewer than that of TR and the proposal, thus the probability of occurrence of out-of-order packets is lower. Consequently, the TCP retransmissions of RR is lower than that of TR. That of the proposal is also lower than that of TR, then the delay equalization between multiple links in the proposal effects the decrease in the occurrence of out-of-order packets, and effects the decrease in TCP retransmissions.

Figures 12(b) and 12(c) show, respectively, the transition of FTP response time and FTP throughput. The FTP response time of SL and the proposal are superior to that of RR and TR. The IP average delay of TR is superior to that of SL, however, the FTP response time of TR is inferior to that of SL. The inversion is caused by the large number of TCP retransmissions in

TR, and the packet distribution of TR lowers the FTP performance. The cause of the inferior FTP response time of RR to that of SL is not the TCP retransmissions, but is the small amount of TCP flow based on TCP congestion window size, then the packet distribution in RR distributes the large number of packets to 16-link, which is narrow bandwidth, and originally lowers IP performance. The number of TCP retransmissions and the FTP response time of the proposal is the same as those of SL. As the above mentioned, the cause of TCP retransmission in SL is the packet loss in 11b-link, but the cause of that in the proposal is the out-of-order packet, that is, the proposal offsets the improvement of IP performance against the out-of-order packets, and does not improve the FTP performance, but does not lower it.

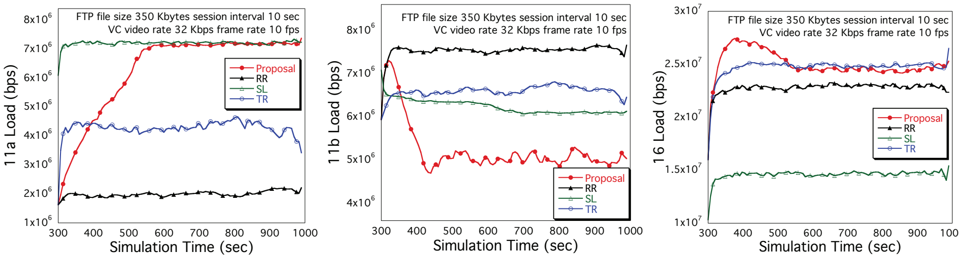
Figures 13(a) and 13(b) show, respectively, the transition of VC average delay and VC throughput. The VC average delay of SL is equal to the IP average delay because a VC frame corresponds to a IP packet and because out-of-order packet does not occur. In the proposal, RR, and TR, the VC average delay is larger than that of IP because the sequence control in VC waits for frame with the expected sequence on the occurrence of out-of-order packet. Therefore, VC average delay of TR is higher than that of SL though IP average delay of TR is lower than that of SL, i.e., the packet distribution of TR lowers the VC performance. On the other hand, that of the proposal is lower than that of SL, therefore, the effect of the packet distribution in the proposal overcomes the ill of it, and can improve the VC performance. That of RR is higher than that of the others because RR originally lowers IP performance.

5.3 Transition of delay and throughput in high traffic load



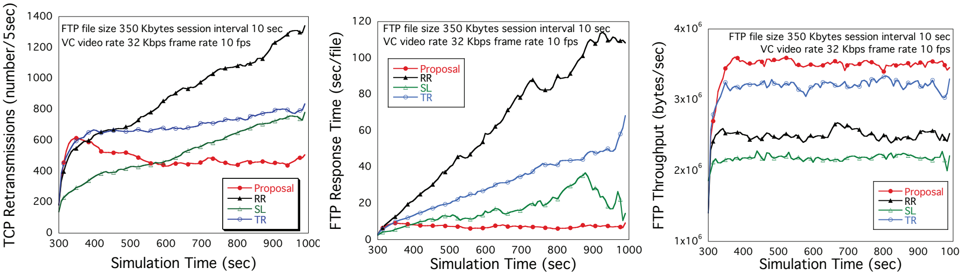
(a) Average delay. (b) Throughput. (c) Out-of-oder.

Fig. 14. Transition of IP on FTP file size 350K bytes.



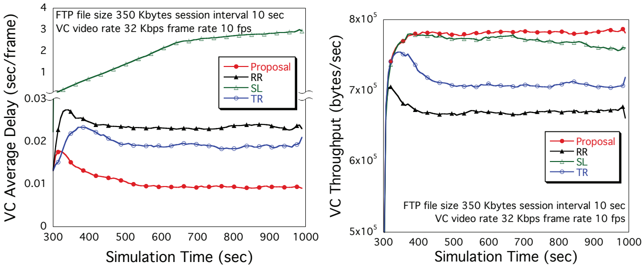
(a) Average delay. (b) Throughput. (c) Out-of-oder.

Fig. 15. Distributed traffic load to each wireless system on FTP file size 350K bytes.



(a) TCP retransmissions. (b) FTP response time. (c) FTP throughput.

Fig. 16. Transition of TCP and FTP on FTP file size 350K bytes.



(a) Average delay. (b) Throughput.

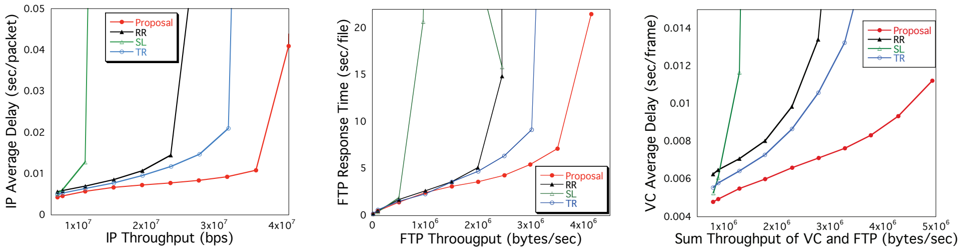
Fig. 17. Transition of VC on FTP file size 350K bytes.

Figures 14(a) and 14(b) show, respectively, the transition of IP average delay and IP throughput, when file size in FTP is 350K bytes, furthermore, Fig. 15(a), 15(b) and 15(c) show, respectively, the transition of 11a load, 11b load and 16 load, when file size in FTP is 350K bytes. The IP average delay of the proposal is low, and is stable. On the other hand, that of the others increase as linear, and become much higher than that of the proposal. Furthermore, their IP throughput are lower than that of the proposal. In area-A, the packet distribute to 11a-link brings low delay to IP because of wide bandwidth and few accommodated terminals in 11a, as mentioned in 5.2. In area-B, the packet collision and loss in 11b further increase because of the increase in traffic, and the large number of retransmissions in MAC brings the increase in delay to IP. Furthermore, the packet loss in 11b brings the decrease in throughput to IP. Each 16-link has the narrow bandwidth, but does not cause the collision because of TDD. i.e., The delay of 16-link is lower than that of 11b-link because of no retransmission process in MAC, which of delay in 11b exponentially increases based on a binary back-off mechanism. Therefore, the large number of packet distribute to 11b brings the increase in delay and the decrease in throughput to IP. Consequently, IP average delay of the proposal, which distributes the smaller number of packets to 11b than the others (see Fig. 15(b)), is lowest, and its IP throughput is highest.

Figures 14(c) and 16(a) show, respectively, the transition of IP out-of-order packets and TCP retransmissions, when file size in FTP is 350K bytes. The IP out-of-order packets of the proposal decreases rapidly as the packet distribute proceeds the same as the case that FTP file size is 1K bytes, i.e., the delay equalization between the multiple links in the proposal effects the decrease in IP out-of-order packets. That of RR also decreases, but the decrease in the amount of TCP flow based on TCP congestion window size, which becomes small rapidly by the increase in IP delay of RR, brings it. TCP retransmission is caused by the IP packet loss and IP out-of-order packets. The TCP retransmissions in SL is caused only by IP packet loss, and IP packet loss is caused by the large number of distributed packets to 11b. That of RR, TR and the proposal is caused by IP packet loss and IP out-of-order packets. That of RR is caused largely by IP packet loss, because RR distributes the large number of packets to 11b and IP out-of-order packets decreases by the decrease in TCP flow. Therefore, the trend of TCP retransmissions of RR is similar to that of SL. TR also distributes the large number of packets to 11b, but distributes the larger number of packets than RR to 11a and 16, which of packet loss probability is much lower than 11b, i.e., the TCP retransmissions in TR is caused mainly by out-of-order packets and it reduces the upward trend of TCP retransmissions in comparison with SL and TR. On the other hand, the TCP retransmissions of the proposal is low stable in comparison with the others. The proposal distributes the much smaller number of IP packets than the others to 11b and reduces IP packet loss, furthermore, it equalizes the delay of each link in M-route, thus reduces also IP out-of-order packets. That brings the low and stable retransmissions to TCP.

Figures 16(b) and 16(c) show, respectively, the transition of FTP response time and FTP throughput, when file size in FTP is 350K bytes. The FTP response time of RR and TR increase as linear. In RR and TR, FTP session can not complete in a period of 10 sec, which is FTP session start interval, because the amount of TCP flow is restrained low by the large number of retransmissions. The active FTP session accumulates. Therefore, the access network causes the congestion. In the proposal, FTP session can complete within 10 sec, and the delay not increase and is stable. Furthermore, the throughput reaches the input load 4M bytes/sec. Therefore, the proposal controls avoids the congestion.

5.4 Dependence of delay on throughput



(a) IP. (b) FTP. (c) VC.

Fig. 18. Dependence of delay on throughput.

Figure 18(a), 18(b), and 18(c) shows, respectively, the dependence of IP average delay on IP throughput, the dependence of FTP response time on FTP throughput , and the dependence of VC average delay on VC throughput when FTP file size increases from 1K bytes to 400K bytes. The average delay and throughput are each the averages for 10 topologies in which the antennas and terminals are deployed randomly in the evaluation space.

When the FTP traffic is low, the performance of SL and the proposal is superior to that of RR and TR. In low load, if packets are distributed to a widest band link, that is, if the packet distribution is equalized to that of SL, the performance becomes high. The packet distribution of the proposal becomes equal to that of SL, but that of RR and TR do not. As FTP traffic increases, the 11b-link load of M-route in 11b-coverage and outside 11a-coverage becomes high, then M-route including 11b-link needs to distribute packets to 11a-link or 16-link. SL can not distribute packets of 11b-link to other links, then SL is saturated first by the exhaustion of 11b-link capacity. By the same cause, RR and TR are saturated in FTP file size 300K bytes and 400K bytes respectively. The proposal distributes packets from 11b-link to 16-link and 11a-link, and avoids the saturation until FTP file size exceeds 400K bytes.

Summarizing, in any FTP traffic, the proposal can distribute packets effectively in comparison with other methods, and it produces low delay and hight throughput on both TCP application and UDP application, and simultaneously.

6. Conclusion

In this chapter, the packet distribution characteristics in IEEE802.11-link and that in IEEE802.16-link was respectively shown, and, based on these characteristics, the packet distribution method for access route compositing IEEE802.11/16-links was proposed. Furthermore, its performance through evaluation with IEEE802.11a/b and IEEE802.16 was shown. Consequently, the proposed method was found to have the following effectiveness.

- It can greatly effectively distribute packets to IEEE802.11/16 links according to link load.
- And, it can also reduce out-of-packets caused by distributing packets to multiple links.
- Then, It can decrease delay and can increase throughput on both TCP application and UDP application, and simultaneously.

7. References

Arkin, D. (2004). *My Life*, Arkin Publishing, Arkinson.

Mitoralll, J. & Maguire, G. (1999). Cognitive Radio: Making Software Radios More Personal, *IEEE Pers. Comm.*, Vol. 6, No. 4, pp. 13–14 1999.

Mitoralll, J. (1999). Cognitive Radio for Flixible Multimedia Communications, *Proc. MoMuC99*, pp. 3–10, 1999.

Harada, H. (2005). A Study on a new wireless communications system based on cognitive radio technology, *IEICE Tech. Rep.*, Vol. 105, No. 36, pp. 117–124, 2005.

3GPP TS 22.258. (2005). Service Requirements for the All-IP Network (AIPN); Stage 1, v2.0.0, 2005.

ITU-T. (2006). Y.2021: NGN Release 1, 2006.

Phatak, D. S. & Goff, T. (2002). A novel mechanism for data streaming across multiple IP links for improving throughput and reliability in mobile environments, *Proc. IEEE INFOCOM*, pp. 773–781, 2002.

Snoeren, A. C. (2002). Adaptive Inverse Multiplexing for Wide-Area-Wireless Networks, *Proc. IEEE GlobCom'99*, Vol.3, pp.1665–1672, 1999.

Shrama, P.; Lee, S.; J. Brassil, J. & Shin, K. (2007). Aggregating Bandwidth for Multimode Mobile Collaborative Communities, *IEEE Tans. on MC*, Vol. 6, No. 3, pp. 280–296, 2007.

Chebrolu, K. & Rao, R. (2006). Bandwidth Aggregation for Real-Time Applications in Heterogeneous Wireless Networks, *IEEE Tans. on MC*, Vol. 5, No. 4, pp. 388–403, 2006.

Hsieh, H.; Kim, K. & Sivakumar, R. (2004). An end-to-end approach for transparent mobility across heterogeneous wireless networks, *Mob. Netw. Appl.*, Vol. 9, No. 4, pp. 363–378, 2004.

Zhang, M.; Lai, J.; Krishnamurthy, A.; Peterson, L. & Wang, R. (2004). A Transport Layer Approach for Improving End-to-End Performance and Robustness Using Redundant Paths, *USENIX 2004*, 2004.

D. Gross, D. & C. Harris, C. (1985). Fundamentals of Queueing Theory 2nd ed, *John Wiley & Sons*, 1985.

Little, J. (1961). A Proof of the Queueing Formula $L = \lambda W$", *Opre Res J.*, 18:172–174, 1961.

Bianchi, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function, *IEEE JSAC*, Vol. 18, No. 3, pp. 535–547, 2000.

Carvalho, M. M. & Garcia-Luna-Aceves, J. J. (2003). Delay analysis of IEEE 802.11 in single-hop networks, *Proc. of ICNP*, pp. 146 –155, 2003.

Takizawa, Y.; Taniguchi, N.; Yamanaka, S.; Yamaguchi, A. & Obana, S. (2008). Characteristics of Packet Distribution in Wireless Access Networks Accommodating IEEE802.11 and IEEE802.16, *IPSJ Journal*, Vol. 49, No. 9, pp. 3245–3256, 2008.

Takizawa, Y.; Taniguchi, N.; Yamanaka, S.; Yamaguchi, A. & Obana, S. (2008). Packet Distribution Control for Wireless Access Networks Accommodating IEEE802.11 and IEEE802.16, *IPSJ Journal*, Vol. 49, No. 10, pp. 3576–3587, 2008.

Bertsekas, D. & Gallager, R. (1992). Data Networks, *Prentice Hall*, 1992.

IEEE Std 802.16-2004. (2004). Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

IIEEE Std. 802.16e-2005. (2005). Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access System, 2005.

Takada, J. (2004). Radiowave Propagation for Mobile Satellite Communications, *Tech. Rep. of IEICE*, Vol. 104, No. 671, pp. 13–16, 2004.

3GPP2. (2006). C30-20060823-004A Evaluation methodology V4.0, 2006.

Konishi, S.; Wang, X.; Kitahara, T.; Nakamura, H. & Suzuki, T. (2008). A Study on Ultra Low-Latency Mobile Networks, *Wireless Personal Comm.:An Int. Journal*, Vol. 44, No.1, pp.57–73, 2008.

Nakayo, D. J. and Hossain, E. (2006). A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks, *IEEE Trans Comp.*, Vol. 55, No.11, pp. 1473–1488, 2006.

Cho, D.; Song, J.; Kim, M. and Han, K. (2006). Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network, *Proc. IEEE DFMA 2005.*, 2005.

Lin, L.; Jia, W. and Lu, W. (2007). Performance Analysis of IEEE 802.16 Multicast and Broadcast Polling based Bandwidth Request, *Proc. IEEE WCNC 2007.*, 2007.

Iyengar, R.; Iyer, P. and Biplab Sikdar, B. (2005). Delay Analysis of 802.16 based Last Mile Wireless Networkst, *Proc. IEEE GlobeCom 2005.*, 2005.

He, J.; Guild, K.; Yang, K. and Chen, H. (2007). Modeling Contention Based Bandwidth Request Scheme for IEEE 802.16 Networks, *IEEE Comm. Letter*, Vol. 11, No.8, pp. 698–700, 2007.

Ni, Q.; Xiao, Y.; Turlikov, A. and Jiang, T. (2007). Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMAX Networks, *IEEE Comm. Magazine*, Vol. 4, No.4, pp. 477–486, 2007.

Part 3

Applications and Realizations

14

Wireless Sensor Network: At a Glance

A.K. Dwivedi[1] and O.P. Vyas[2]
*[1]School of Studies in Computer Science & Information Technology,
Pandit Ravishankar Shukla University, Raipur, C.G.,
[2]Indian Institute of Information Technology-Allahabad (IIIT-A),
Deoghat, Jhalwa, Allahabad, U.P.,
India*

1. Introduction

Wireless Sensor Network is a technology which has capability to change many of the Information Communication aspects in the upcoming era. From the last decade Wireless Sensor Networks (WSNs) is gaining magnetic attention by the researchers, academician, industry, military and other ones due to large scope of research, technical growth and nature of applications etc. Wireless Sensor Networks (WSNs) employ a large number of miniature disposable autonomous devices known as sensor nodes to form the network without the aid of any established infrastructure. In a Wireless Sensor Network, the individual nodes are capable of sensing the environments, processing the information locally, or sending it to one or more collection points through a wireless link. Day to day applications of WSNs is increasing from domestic use to military use and from ground to space.

The objective of this book chapter is to explore all aspects of WSNs under different modules including these as well in a systematic flow: Sensor nodes, Existing hardware, Sensor node's operating systems, node deployment options, topologies used for WSN, architectures, WSN lifecycle, Resource constraint nature, Applications, Existing experimental tools, Usability & reliability of experimental tools, Routing challenges and Protocol design issues, Major existing protocols, Protocol classifications, Protocols evaluation factors, Theoretical aspects of major Energy Efficient protocols, Security issues, etc.

This chapter contains from very basic to high level technical issues obtained from highly cited research contribution in a concluding manner but presenting whole aspects related to this field.

2. Wireless sensor nodes and existing hardware

Wireless sensor nodes are tiny, light weight sensing devices consists of a constrained processing unit, little memory, EEPROM or Flash memory for tiny operating systems and other desired programs, one or more sensors, a limited range transceiver, battery or solar based power unit and optionally a mobility subsystem for mobile sensor nodes (Dwivedi & Vyas, 2010).

Tatiana Bokareva presented a mini hardware survey related to wireless sensor nodes (Tatiana), except this a comprehensive listing of existing wireless sensor nodes is presented

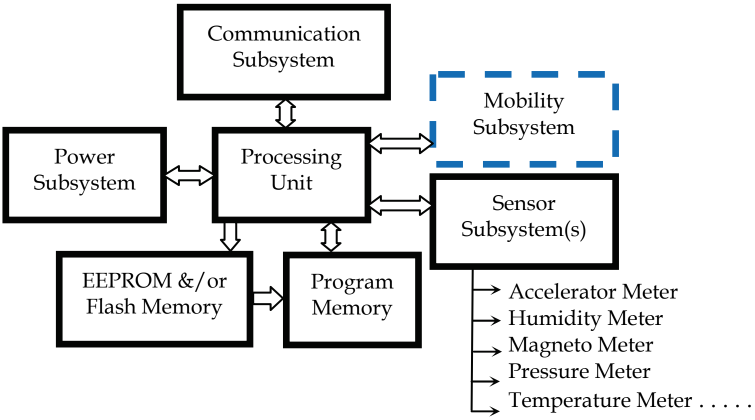


Fig. 1. Block diagram of wireless sensor node

and maintained by Imperial College London (ICL, 2007), Embedded WiSeNts Platform Survey (Embedded WiSeNts, 2006) presents an in-depth survey of five popular wireless sensor nodes (ESB/2, BTnode, uNode, Tmote Sky, and EYES IFXv2), another pretty listing is presented by University of California's Sensor Network Systems Laboratory (Senses, 2005). As well as Sensor Network Museum (SNM, 2010) maintained by TIK computer Engineering and Networks Laboratory, ETH Zurich presents a collection of reference data and links for commonly used wireless sensor nodes and related links. In a research contribution (Manjunath, 2007), technical specifications of some well known wireless sensor nodes are presented in tabular format, as here in its original (Table 1).

Resource footprint (Tatiana; ICL, 2007; Embedded WiSeNts, 2006; Senses, 2005; SNM, 2010; Manjunath, 2007) for various currently available Wireless Sensor nodes provides us a summary that most of the Nodes belongs to within the following configuration:

- 4-bit to 8-bit processor
- 512 Byte to 512 KB RAM (Program and Data Memory)
- 4 KB to 4 MB Flash/External Memory
- 250 Kbps 2.4 GHz IEEE 802.15.4 or Bluetooth 2.0 or 10 Kbps etc. as radio transceiver

On the basis of above mentioned resource footprint it can be concluded that each and every currently available sensor nodes face limited resource problems such as narrow address space and slow clock cycle of micro controller, small program and data memory as well as external memory, low bandwidth and low range of transceivers.

Table 2 presents a wider look on technical aspects of some hardware systems for WSNs, because hardware designing requires a holistic approach for WSNs, looking at all areas of the design space. Expanding the uses of WSNs for various applications, expect more performance for less power out of the hardware platforms. Envision a future of WSNs made up of ultra low power nodes that provide high power computation and can be deployed for decades is possible only with more research effort (Hempstead et al., 2008).

3. Operating systems for wireless sensor nodes

WSNs are composed of large numbers of tiny-networked devices that communicate untethered. Operating systems are at the heart of the sensor node architecture. In terms of

S.N.	Platform	MCU	RAM	Code Memory	RF Transceiver	Frequency	Radio range (feet)
1.	Mica	Atmel ATMega128L	4KB	128KB	TR1000	433, 916 MHz	200
2.	Mica2	Atmel ATMega128L	4KB	128KB	CC1000	315, 433, 916 MHz	500
3.	Mica2Dot	Atmel ATMega128L	4KB	128KB	CC1000	315, 433, 916 MHz	500
4.	MicaZ	Atmel ATMega128L	4KB	128KB	CC2420	2.4 GHz	410
5.	Cricket	Atmel ATMega128L	4KB	128KB	CC1000	433 MHz	500
6.	TelosA	TIMSP430	2KB	60KB	CC2420	2.4 GHz	410
7.	TelosB	TIMSP430	10KB	48KB	CC2420	2.4 GHz	410
8.	BTnode3	Atmel ATMega128L	64KB	128KB	Zeevo-BT/CC1000	2.4 GHz/868 MHz	328/500
9.	EYES	TIMSP430	4KB	60KB	TR1001	868 MHz	984
10.	Intel mote	ARM7TDMI (Core)	64KB	512KB	Zeevo-BT	2.4 GHz	328
11.	Intel mote2	PXA27x (Core)	256KB	32MB	CC2420	2.4 GHz	410
12.	MANTIS nymph	Atmel ATMega128L	4KB	128KB	CC1000	315, 433, 868, 915 MHz	500
13.	XYZ mote	ARM7TDMI (Core)	32KB	256KB	CC2420	2.4 GHz	410
14.	ECR	TIMSP430	2KB	60KB	TR1001	868 MHz	984
15.	ESB	TIMSP430	2KB	60KB	TR1001	868 MHz	984
16.	Smart-Its mote	Atmel ATMega103L	4KB	128KB	Ericsson-BT/TR1001	2.4 GHz/ 868 MHz	328/984
17.	Tmote Sky	TIMSP430	10KB	48KB	CC2420	2.4 GHz	410
18.	TinyNode 584	TIMSP430	10KB	48KB	Xemics XE1205	868 MHz	200
19.	ZebraNet H/W	TIMSP430	2KB	60KB	9XStream	902-928 MHz	328

Table 1. A summarized list of some popular wireless sensor node (Manjunath, 2007)

SN	System	Arch Style	Data path width	Event driven (Y/N)	Circuit Techniques	Accelerators	Memory (KB)	Process	Voltage (V)	Throughput (MIPS)	Energy (pJ/ins)
1.	Atmel ATmega128L	GP Off-the-shelf	8	N	N	N	132KB	350nm	3.0	7.3 MHz	3200
2.	TI MSP430	GP Off-the-shelf	16	N	N	N	10KB	NA	3.0	8 MHz	750
3.	SNAP/LE	GP RISC	16	Y	Asynchronous	Timer, message interface	8KB	180nm	1.8 0.6	200 23	218 24
4.	BitSNAP	GP RISC Bit-serial datapath	16	Y	Asynchronous	Timer, message interface	8KB	180nm	1.8 0.6	54 6	152 17
5.	Smart Dust	GP RISC	8	N	Synchronous - 2 clock	None	3.125KB	250nm	1.0	0.5 (500KHz)	12
6.	Charm	Protocol processor	NA	N	Two power domains	Custom radio stack	68KB	130nm	1.0V (high) 0.3-1.0V (low)	8 MHz	150μW 53.6 μW leakage
7.	Michigan 1	GP	8	Y	Sub-threshold	None	0.25KB	130nm	0.360	833 KHz	2.6
8.	Michigan 2	GP	8	Y	Sub-threshold	None	0.3125KB	130nm	0.350	354 KHz	3.52
9.	Harvard	Event driven accelerator	8	Y	VDD-gating	Timer, filter, message proc	4KB	130nm	0.55-1.2	12.5 MHz	680 pJ/task

Table 2. Technical specification for some hardware systems for Wireless Sensor Network (Hempstead et al., 2008)

Wireless Sensor Networks we need these things in operating system architectures: Extremely small footprint, extremely low system overhead and extremely low power consumption. When designing or selecting operating systems for tiny-networked sensors, our goal is to strip down memory size and system overhead because typical wireless sensor nodes are equipped with a constrained processing unit, little memory, EEPROM or Flash memory, battery or solar based power unit. In a research contribution (Hempstead et al., 2008) and in a technical report (Fröhlich & Wanner, 2008) three classifications of O. S. architectures are described for wireless sensor nodes: Monolithic, Modular/Micro and Virtual Machine.

After evaluating various research contributions specifically devoted to operating systems used for wireless sensor nodes (Fröhlich & Wanner, 2008, Reddy et al., 2007; Dwivedi et al., 2009a; Manjunath, 2007) total 39 operating systems are identified:

1.	TinyOS	2.	Contiki	3.	Mantis OS
4.	Microsoft .NET Micro	5.	YATOS (Yet Another Tiny OS)	6.	BTnutOS or NutOS
7.	PeerOS	8.	Embedded Linux	9.	NanoRK
10.	µCOS	11.	Squawk VM	12.	SensorOS
13.	MagnetOS	14.	CORMOS	15.	Bertha
16.	kOS	17.	VMSTAR	18.	Maté
19.	CVM	20.	EYES	21.	SenOS
22.	DCOS	23.	t-Kernel	24.	Nano-QPlus
25.	SmartOS	26.	AVRX	27.	Pixie
28.	LiteOS	29.	T2	30.	OSSTAR
31.	Jallad	32.	CustomOS	33.	GenOS
34.	MoteWorks	35.	NanoVM	36.	ParticleVM
37.	KVM	38.	AmbiCompVM	39.	SOS

Table 3. List of operating systems available for Wireless Sensor Nodes

D. Manjunath presents a review of current operating systems for WSNs (Manjunath, 2007) whose aims were to explicate "why sensor operating systems are designed the way they are". This technical report questions every design decision, and provide a detail reasoning for why these decisions.

4. Node deployment options in wireless sensor networks

As we know that WSN is deployed to measure environment parameters in Region of Interest (ROI) and to send it to a controller node or base station. In WSNs how nodes will deployed is basically application specific and totally dependent on environment. The node deployment option affects the performance of routing protocol basically in terms of energy consumptions. Basically there are three ways in which tiny sensor nodes can be deployed in a wireless sensor network environment:

- **Regular Deployment** - Sensor nodes can be deployed in a well planned, fixed manner; not necessarily geometric structure, but that is often a convenient assumption. In this type of deployment data is routed through a predefined path.
  Area of Use: Medical and health, Industrial sector, Home networks, etc.

- **Random Deployment** – Sensor nodes are scattered over finite area. When the deployment of nodes is not predefined optimal positioning of cluster head becomes a critical issue to enable energy efficient network operation. Random deployment is generally used in rescue operations.
  Area of Use: Environmental and Habitual monitoring, etc.
- **Sensor Nodes with Mobility** - Can move to compensate for deployment shortcomings; can be passively moved around by some external force (wind, water, vehicle); can actively seek out "interesting" areas.
  Area of Use: Battle field surveillances, Emergency situations (Fire, Volcano, Tsunami), etc.

5. Topologies used for wireless sensor networks

Wireless sensor nodes are typically organized in one of three types of network topologies:

- In a **star topology**, each node connects directly to a gateway.
- In a **cluster tree topology**, each node connects to a node higher in the tree and then to the gateway, and data is routed from the lowest node on the tree to the gateway.
- Finally, to offer increased reliability, **mesh networks** feature nodes that can connect to multiple nodes in the system and pass data through the most reliable path available.

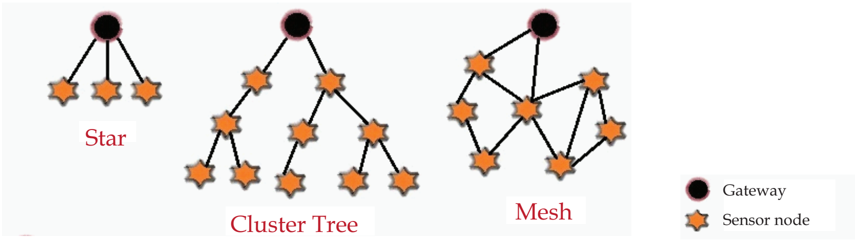


Fig. 2. Topologies used for Wireless Sensor Networks

Three phases related to topology maintenance and changes has been presented in a research contribution (Akyildiz et al., 2002a):

- Pre-deployment and Deployment phase
- Post-deployment phase
- Redeployment of additional nodes phase

6. Architectures for wireless sensor networks

In a technical report (Karl & Willig, 2003) Holger Karl and Andreas Willig present views on WSN architectures in the light of principle differences in application scenarios and underlying communication technology. The architecture of WSNs will be drastically different both concerning a single node and the network as a whole. Wide range of sensor node architectures has been presented till today but as a general design principle all of them have targeted the following objectives: energy efficiency, small size and low cost. The architecture for network as a whole is a set of principles that guide where functionality should be implemented along with a set of interfaces, functional units, protocols, and physical hardware that follows those guidelines.

In another research paper (Dulman & Havinga, 2005) the characteristics of wireless sensor networks from an architectural point of view is presented. Since WSNs are designed for specific applications so there is no precise architecture to fit them all but rather a common set of characteristics that can be taken as a starting point. In same paper a data-centric architecture is also presented.

A research paper (NeTS-NOSS, 2007) presents six aspects of architecture for WSN: Design Principles, Functional Architecture, Programming Architecture, Protocol Architecture, System Support Architecture and Physical Architecture. This paper also states that "The situation today in sensor networks is that none of these six levels of network system architecture are 'solved' or even clearly established. The vast majority of the studies fall in the category of protocol architecture".

In a research paper (Vazquez et al., 2009), an architecture for integrating Wireless Sensor Networks into the Internet of Things called "Flexeo" is presented. In another research paper (Schott et al., 2007) a flexible protocol architecture "e-SENSE" for WSNs has been introduced, which is well-suited for capturing the context surrounding service users in order to enable a variety of advanced context-aware applications in beyond 3G mobile communication systems.

7. Wireless sensor networks lifecycle

Characteristically, there are four phases in the lifecycle of a wireless sensor network (the implementation phase is omitted because the sensor code is frequently reused). Researchers are usually involved in the planning and deployment phase, while the final customers are more interested in monitoring and control the WSN.

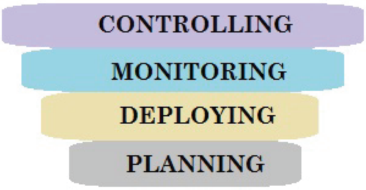


Fig. 3. Wireless sensor network lifecycle

Planning WSNs Planning phase usually involves the inspection of the deployment area and the selection of the correct locations to position the sensors in a way that accomplishes the intended goal.

Deploying WSNs In the deployment phase, sensor nodes continually send their wireless connection quality and route to the base.

Monitoring WSNs In this phase, the user interest is mainly focuses on the values read by network sensors.

Controlling WSNs The application can also be used to control WSNs by sending commands to the network. These commands can tell the network devices to stop sending messages, increase the time between messages or even reset the network (restart the Multi-Hop algorithm). In future, WSNs could be controlled via a web interface or a handheld device, being easier to stop and restart the network as needed.

8. Resource constraint nature of wireless sensor networks

Wireless Sensor Networks (WSNs) employ a large number of miniature disposable autonomous devices known as sensor nodes to form the network without the aid of any established infrastructure. In a Wireless Sensor Network, the individual nodes are capable of sensing their environments, processing the information locally, or sending it to one or more collection points through a wireless link. Sensor node may fail due to lack of energy, physical damage, communications problem, inactivity (a node becomes suspended), or environmental interference. Resource footprint for various currently available Wireless Sensor nodes is presented in section 2, obtained from (Tatiana; ICL, 2007; Embedded WiSeNts, 2006; Senses, 2005; SNM, 2010; Manjunath, 2007). Here is a table focuses on resource constraint nature of Wireless Sensor Nodes and obviously WSNs:

Node	**CPU**	**Memory**	**Radio**
Rene 1999	ATMEL 8535	512 Byte RAM 8 KB Flash	10 Kbps
Mica-2 2001	ATMEGA 128	4 KB RAM 128 KB Flash	76 Kbps
Telos 2004	Motorola HC 508	4 KB RAM	250 Kbps
Mica-Z 2004	ATMEGA 128	4 KB RAM 128 KB Flash	250 Kbps
BT Node 2001	ATMEL Mega 128L	128 KB Flash 4 KB EEPROM 4 KB SRAM	Bluetooth
Imote 1.0 2003	ARM 7TDMI	64 KB SRAM 512 KB Flash	Bluetooth
Stargate 2003	Intel PXA 255	64 KB SRAM	Serial Connection to Sensor Network
Insysnc Cerfoube 2003	Intel PXA 255	32 KB Flash 64 KB SRAM	
PC 104	X86 Processor	32 KB Flash 64 KB SRAM	

Table 4. Presenting resource constraint nature of some popular wireless sensor nodes

9. Applications of wireless sensor networks

WSNs can be applied in a wide range of areas, such as: habitat monitoring and tracking, disaster relief, emergency rescue operation, home networks, detecting chemical/biological/radiological/nuclear/explosive material, monitoring patents and elderly people, asset and warehouse management, building monitoring and control, fleet monitoring, military battlefield awareness and surveillance, security and surveillance, environmental monitoring, pipeline corrosion monitoring, homeland security, monitoring conditions of buildings and bridges, industrial process monitoring and control, machine health monitoring, healthcare applications, home automation, traffic control, etc.

With the help of research contributions (Biradar et al., 2009; Katiyar et al., 2010) a table is presented here, which systematically summarized some applications for different areas:

Area	**Applications**
Military	• Military situation awareness. • Sensing intruders on basis. • Detection of enemy unit movements on land and sea. • Battle field surveillances
Emergency situations	• Disaster management. • Fire/water detectors. • Hazardous chemical level and fires.
Physical world	• Environmental monitoring of water and soil. • Habitual monitoring. • Observation of biological and artificial systems. • Marginal Farming.
Medical and health	• Sensors for blood flow, respiratory rate, ECG (electrocardiogram), pulse oxymeter, blood pressure and oxygen measurement. • Monitoring people's location and health condition.
Industry	• Factory process control and industrial automation. • Monitoring and control of industrial equipment. • Machine health monitoring.
Home networks	• Home appliances, location awareness (blue tooth). • Person locator.
Automotive	• Tire pressure monitoring. • Active mobility. • Coordinated vehicle tracking.
Area monitoring	• Detecting enemy intrusion • Geo-fencing of gas or oil pipelines. • Detecting the presence of vehicles.
Environmental monitoring	• Air pollution monitoring. • Forest fires detection. • Greenhouse monitoring. • Landslide detection. • Volcano monitoring. • Flood detection.
Water/Wastewater monitoring	• Landfill ground well level monitoring and pump counter. • Groundwater arsenic contamination assessment. • Measuring water quality.
Cognitive sensing	• Bio-inspired sensing. • Swarm intelligence. • Quorum sensing.
Underwater acoustic sensor systems	• Oceanographic data collection. • Pollution monitoring. • Disaster prevention. • Assisted navigation. • Tactical surveillance.
Traffic Management & Monitoring	• Traffic congestion control. • Road Surface Condition Monitoring (BusNet in Sri Lanka).

Table 5. Some applications of WSNs in different areas

Deploying nodes in an unattended environment will provide more possibilities for the exploration of new applications. WSNs will be ubiquitous in the near future, due to new opportunities for the interaction between humans and their physical world also WSNs are expected to contribute significantly to pervasive computing.

10. Existing standards for wireless sensor networks

WSNs fascinate a number of standardization bodies to develop standards, due to a smaller amount of standards exists for WSNs in comparison to other wireless networks. A number of standards are currently under development or ratified for WSNs. Some standardization bodies working in the specific field of WSNs to setup standards, such as:

Standardization body	**Specific work area for WSN**
Institute of Electrical and Electronics Engineers	Physical layer and MAC sub layer of Data link layer.
Internet Engineering Task Force	Data link layer and all above layers of WSN protocol stack.
International Society of Automation	All layers of WSN protocol stack
DASH7 Alliance	Promotes the use of the ISO 18000-7 standard for wireless sensor networks.

Table 6. Some main Standardization bodies and their specific work area

Apart from these several non-standard, proprietary mechanisms and specifications also exist. The most commonly used predominant standards in WSNs include:

IEEE 802.15.4	Standard for low-rate, wireless personal area networks, defines the "physical layer" and the "medium access layer".
Zigbee	ZigBee builds upon the 802.15.4 standard to define application profiles that can be shared among different manufacturers.
IEEE 802.11	Standards efforts for low-power Wi-Fi.
IEEE 1451	The objective of this standard is to make it easier for different manufacturers to develop smart sensors and to interface those devices to networks.
ISA100	Addresses wireless manufacturing and control systems in the areas of the: Environment, Technology and life cycle, and Application of Wireless technology.
6LoWPAN	IPv6 over low-power wireless networks, defines an adaptation layer for sending IPv6 packets over IEEE 802.15.4.
uIPv6	uIPv6 is the world's smallest certified open source IPv6 stack provides TCP/IP connectivity to tiny embedded 8-bit micro controllers for low-cost networked device such as sensors and actuators with maintained interoperability and RFC standards compliance.

Table 7. Predominant standards in field of WSNs

11. Existing experimental tools for wireless sensor networks

Research activities in the area of Wireless Sensor Networks (WSNs) need expositive performance statistics about scenario, systems, protocols, gathered data, applications and many more. There are various experimental tools for fulfilling these requirements, someone are in practical use while other one are in literatures. In this part of chapter a glance on currently available simulation tools/frameworks, emulators, visualization tools, testbeds, debuggers, code-updaters and network monitoring tools used for wireless sensor networks is presented (Dwivedi & Vyas, 2011).

11.1 Simulator/simulation framework

A simulator is a software that imitates selected parts of the behavior of the real world. Depending on the intended usage of the simulator, different parts of the real-world system are modeled and imitated. The parts that are modeled can also be of varying abstraction level. A wireless sensor network simulator imitates the wireless media and the constraints nodes in the network. Some sensor network simulators have a detailed model of the wireless media including effects of obstacles between nodes, while other simulators have a more abstract model.

Type of simulation

Simulators either run as in an asynchronous mode, event triggered mode, or in synchronous mode, where events happen in parallel in fixed time slots (DCG's Sinalgo, 2009):

- *Synchronous simulation*
  The synchronous simulation is purely based on rounds.
- *Asynchronous Simulation*
  The asynchronous simulation is purely event based.

Categorization of simulators

A large number of sensor network simulators have been proposed by researchers. In a research contribution WSN Simulators are categorized (Eriksson, 2009) as:

- *Generic Network Simulators*
- *Code Level Simulators*
- *Firmware Level Simulators*

In another research contribution (Shu et al., 2009), simulators have been classified into the following three major categories based on complexity:

- *Algorithm Level Simulators*
- *Packet Level Simulators*
- *Instruction Level Simulators*

Several simulators exist that are either adjusted or developed specifically for wireless sensor networks. Here is a table presenting **63** simulators/simulation frameworks (Table 8).

11.2 Emulator or emulation environment

As a networked embedded system, a WSN application involves sensor node hardware, its drivers, operating systems, and networking protocols. As a result, the performance of the WSN application depends on all of these factors in addition to its implementation. An emulator is a special type of simulator whose aims is to enable realistic performance evaluation for WSN applications. Emulation environment or emulators are good choice, in

1.	Network Simulator (NS)	2.	Mannasim (NS-2 Extension for WSNs)	3.	DiSenS (Distributed SENsor network Simulation)
4.	(J) Prowler	5.	LecsSim	6.	WISDOM
7.	TOSSIM	8.	OPNET	9.	Sinalgo
10.	TOSSF	11.	SENS	12.	SENSORIA
13.	PowerTOSSIMz	14.	EmStar/Em*	15.	Capricorn
16.	ATEMU	17.	EmTOS	18.	SIDnet-SWANS
19.	COOJA	20.	SenQ	21.	Stargate Simulator (starsim)
22.	GloMoSim (Global Mobile Information Systems Simulation)	23.	H-MAS (Heterogeneous Mobile Ad-hoc Sensor-Network Simulation Environment)	24.	JiST/SWANS (Java in Simulation Time/ Scalable Wireless Ad hoc Network Simulator)
25.	QualNet	26.	SensorSim	27.	SNSim
28.	SENSE	29.	Shawn	30.	SNIPER-WSNSim
31.	VisualSENSE	32.	NetTopo	33.	SNAP
34.	AlgoSenSim	35.	Atarraya	36.	SimPy
37.	Georgia Tech Network Simulator (GTNetS)	38.	SSFNet (Scalable Simulation Framework)	39.	Mule
40.	OMNet++	41.	WiseNet	42.	CaVi
43.	Castalia	44.	SimGate	45.	Ptolemy
46.	J-Sim (formerly JavaSim)	47.	SimSync	48.	Maple
49.	Mote simulator (motesim)	50.	SNetSim	51.	WISENES (WIreless SEnsor NEtwork Simulator)
52.	JiST/SWANS++	53.	SensorMaker	54.	WSNet-Worldsens and WSim
55.	Avrora	56.	TRMSim-WSN	57.	LSU SensorSimulator
58.	Sidh	59.	*PAWiS*	60.	WSNGE
61.	Prowler	62.	OLIMPO	63.	TikTak

Table 8. Simulator/simulation frameworks specifically designed for WSNs

which WSN applications can be directly run for testing, debugging, and performance evaluation. Additionally, studies on the lower layers (e.g., hardware drivers, OS, and networking) as well as cross-layer techniques can also be done in this environment by plugging the target modules into the emulator. Here is a table which presents **14** emulators:

1.	VMNET	2.	Freemote	3.	UbiSec&Sens
4.	ATEMU	5.	EmPro	6.	Emuli
7.	Emstar	8.	NetTopo	9.	MSPSim
10.	TOSSIM	11.	OCTAVEX	12.	MEADOWS
13.	AvroraZ/Avrora	14.	SENSE		

Table 9. Emulators specifically designed for WSNs

11.3 WSN data visualization tools

With the increase in applications for sensor networks, data manipulation and representation have become a crucial component of sensor networks. The data gathered from WSNs is usually saved in the form of numerical form in a central base station. There are many programs that facilitate the viewing of these large amounts of data. These special programs are called data visualization tool for WSNs. Visualization tools can support different data types, and visualize the information using a flexible multi-layer mechanism that renders the information on a visual canvas. Here is a table presenting **19** data visualization tools (Parbat et al., 2010) that are especially designed and developed for WSNs applications:

1.	SpyGlass	2.	TOSGUI	3.	*Oscilloscope*
4.	MoteView	5.	MSR Sense	6.	GSN
7.	TinyViz	8.	Trawler	9.	WiseObserver
10.	XbowNet	11.	SNAMP	12.	SenseView
13.	MonSense	14.	Surge Network Viewer	15.	*MeshNetics WSN Monitor*
16.	NetTopo	17.	Mica Graph Viewer	18.	MARWIS
19.	Octopus				

Table 10. Data visualization tools specifically designed for WSNs

1.	Motelab	2.	NetEye	3.	Sharesense
4.	NESC-Testbed	5.	INDRIYA	6.	Trio
7.	WUSTL	8.	CLARITY	9.	sMote
10.	CitySense	11.	GNOMES	12.	CTI-WSN Testbed
13.	Kansei	14.	WSNTB	15.	FEEIT WSN Testbed
16.	MISTLAB	17.	TWIST	18.	Roulette
19.	Orbitlab	20.	X-sensor	21.	BigNet
22.	Emulab	23.	ENL Sensor Network Testbed	24.	UCR Wireless Networking Research Testbed
25.	WISEBED (Wireless Sensor Network Testbeds)	26.	Imote2 Sensor Network Testbed	27.	SWOON (Secure Wireless Overlay Observation Network)
28.	REALnet	29.	PICSENSE	30.	WHYNET
31.	KonTest	32.	SOWNet	33.	CENS-Testbed
34.	SANDbed	35.	IP-WSN Testbed	36.	SCADDS WSN Testbeds
37.	BANAID	38.	SenseNet	39.	Crossbow WSN Testbed
40.	Motescope	41.	Omega	42.	GaTech Testbed
43.	Tutornet: A Tiered Wireless Sensor Network Testbed	44.	CENSE (A Century of Sensor nodes)	45.	Intel Research Berkeley's 150-mote SensorNet Testbed
46.	WINTeR (Wireless Industrial Sensor Network Testbed for Radio-Harsh Environments)	47.	FireSenseTB: A wireless sensor networks testbed for forest fire detection (Kosucu et al., 2009)		

Table 11. Testbeds used for experimental usage specifically for WSNs

11.4 Testbeds for WSN

To achieve high-fidelity in WSN experiments use of testbed is very productive. Testbeds are an environment that provides support to measure number of physical parameters in controlled and reliable environment. This environment contains the hardware, instrumentations, simulators, various software and other support elements needed to conduct a test. Generally, testbeds allow for rigorous, transparent and replicable testing. By providing the realistic environments for testing the experiments, the testbeds bridge the gap between the simulation and deployment of real devices. The testbeds thus deployed can improve the speed of innovation and productive research. Here is a table presenting **47** testbeds, used for experimental purposes in various universities, colleges, research institutions or by individuals (Table 11).

11.5 Debugging tools/services/concepts

Due to extreme resource constraints nature, deployment in harsh and unattended environments, lack of run-time support tools and limited visibility into the root causes of system and application level faults make WSNs notoriously difficult to debug. Currently, most debugging systems in WSNs are aimed at diagnosing specific faults, such as detection of crashed nodes, sensor faults, or identifying faulty behavior in nodes. There are few debugging solutions for WSNs available, with a fairly wide range of goals and feature sets. Debuggers for WSNs have been categorized (Tavakoli, 2007) into three distinct categories: source-level debuggers, query-oriented debuggers, and decision-tree debuggers. Here is a table presenting **26** debuggers/debugging concepts/debugging concepts:

1.	Clairvoyant	2.	S$_2$DB	3.	ActorNet
4.	Dustminer	5.	Envirolog	6.	ANDES
7.	Sympathy	8.	NodeMD	9.	EvAnT
10.	FIND	11.	StackGaurd	12.	KleeNet
13.	Passive Distributed Assertions (PDA)	14.	Storage-centric method for Debugging	15.	Model-based diagnosis for WSNs
16.	Chowkidar	17.	Marionette	18.	Post-Deployment Performance Debugging (PD2)
19.	Nucleus-NMS	20.	REDFLAG	21.	Declarative Tracepoints
22.	Debugging WSNs Using Mobile Actors	23.	Monitored External Global State (MEGS)	24.	SNTS: Sensor Network Troubleshooting Suite
25.	Wringer	26.	MDB		

Table 12. Debugging tools/services/concepts specifically useful for WSNs

11.6 Code-updation/reprogramming tool

Large scale WSNs may be deployed for long periods of time during which the requirements from the network or the environment in which the nodes are deployed may change. This may necessitate modifying the executing application or re-tasking the existing application with different sets of parameters, which will collectively refer to as code-updation/reprogramming. The relevant forms of code-updation/reprogramming are (Panta et al., 2009):

- *Remote Multi-hop Reprogramming*
- *Incremental Reprogramming*

Incremental Reprogramming poses several challenges. A class of operating systems, including the widely used TinyOS, does not support dynamic linking of software components on a node. SOS and Contiki, do support dynamic linking, however, their reprogramming support also does not handle changes to the kernel modules. Here is a table presenting **10** code-updaters/reprogramming (Table 13).

1.	Trickle	2.	Deluge	3.	Hermes
4.	FlexCup	5.	Stream	6.	FIGARO
7.	Zephyr	8.	MNP (Multi-hop network reprogramming)	9.	Multihop Over-the-Air Programming (MOAP)
10.	MARWIS (Management ARchitecture for WIreless Sensor Networks)				

Table 13. Code-updaters/Reprogramming tools specifically designed for WSNs

11.7 Network monitoring tools

WSNs are typically composed of low cost tiny hardware devices and tend to be unreliable, with failures a common phenomenon. Accurate knowledge of network health status, including nodes and links of each type, is critical for correctly configuring applications on really deployed WSN and/or WSN testbeds and for interpreting the data collected from them. Here is a table presenting **8** networks monitoring:

1.	Memento	2.	Sympathy	3.	LiveNet
4.	NUCLEUS	5.	HERMES	6.	Chowkidar
7.	DiMo	8.	MARWIS (Management Architecture for heterogeneous Wireless Sensor Networks)		

Table 14. Network monitoring tools specifically designed for WSNs

12. Usability & reliability of experimental tools

The statistics gathered from experimental tools can be realistic and convenient, but due to cost of large number of sensors most researches in wireless sensor networks area is performed by using these experimental tools in various universities, institutes, and research centers before implementing real one. These experimental tools provide the better option for studying the behavior of WSNs before and after implementing the physical one.

Simulators are commonly used for rapid prototyping and also used for the evaluation of new network protocols and algorithms as well as enable repeatability because they are independent of the physical world and its impact on the objects. Moreover, simulations enable nonintrusive debugging at the desired level of detail. In a research contribution various factors have been presented that influences simulation results (Dwivedi et al., 2010).

For successful WSN development cooperation not only between test-beds and simulators but also between simulators is required, however, simulators are usually not designed with cooperation in mind (Li et al., 2010).

13. Routing challenges & protocol design issues in WSNs

Routing in WSNs is very challenging due to unique inherent characteristics (energy efficiency and awareness, connection maintenance, minimum resource usage limitation, low latency, load balancing in terms of energy used by sensor nodes, etc.) that distinguish this network from other wireless networks such as mobile ad hoc networks, cellular networks,

SN	Main Category	Sub Categories
1.	Classification based on Network Structure (Al-Karaki & Kamal, 2004)	• *Flat-based or Data Centric routing:* In flat-based routing algorithm, all nodes play the same role and mainly apply flood based data transferring. • *Hierarchical-based or Cluster based routing:* Hierarchical protocols aim at clustering the nodes so that cluster heads can do some aggregation and reduction of data in order to save energy. Hierarchical routing is mainly two-layer routing where one layer is used to select cluster heads and other for routing. • *Location-based routing:* Location-based protocols utilize the position information to relay the data to the desired regions rather than the whole network.
2.	Classification based on Protocol Operation (Al-Karaki & Kamal, 2004)	• *Multipath-based routing:* This type of routing protocols uses multiple paths instead of a single path in order to enhance network performance. • *Query-based routing:* In this type of routing protocol destination nodes propagate a query for data (sensing task) from a node through the network, and a node with this data sends the data that matches the query back to the node that initiated the query. • *Negotiation-based routing:* These protocols use high-level data descriptors in order to eliminate redundant data transmissions through negotiation. Communication decisions are also made based on the resources available to them. • *QoS-based routing:* In QoS-based routing protocols, the network has to balance between energy consumption and data quality. In particular, the network has to satisfy certain QoS metrics (delay, energy, bandwidth, etc.) when delivering data to the base station. • *Non-coherent & Coherent data-processing based routing:* In non-coherent data processing routing, nodes will locally process the raw data before it is sent to other nodes for further processing.
3.	Classification based on Packet Destination (Karl & Willig, 2006)	• *Gossiping and agent-based unicast forwarding:* These schemas are an attempt of working without routing tables in order to minimize the overflow needed to build the tables, as much as result of the initial stages in which the tables were not built yet. • *Energy-efficient unicast:* These techniques analyze the network nodes distribution to set the cost of transmitting over the link between two nodes and select an algorithm to calculate the minimum cost. • *Broadcast and multicast:* Many nodes must collect or distribute the information to every node in the network (broadcast). In a similar way, sometimes it is necessary to distribute data to a subset of previously known nodes. This process is called multicast. • *Geographic routing:* This kind of routing appeared due to two main motivations: (1) sending data randomly to every node in a given region is called geo-casting; (2) the destination node location must be specified geographically or relatively (with a location service). • *Mobile nodes:* These aspects with motion ability should be considered for wireless sensor networks: mobile sensor nodes, mobile base station, mobile sensed phenomenon or combination of these.
4.	Crossbow (Xbow) classification (Olivares et al., 2007)	• *Basic routing (with normal or improved variants)* • *Reliable routing* • *Low Power routing* • *XMesh routing*

Table 15. Protocol classifications and sub-classifications for WSNs

and wireless mesh networks. Major Constraints while designing protocols for WSNs are: Energy, Processing power, Memory. In various literatures or research contributions, related

SN	Main Category	Sub Categories
5.	Classification based on State (Eriksson, 2009)	• *Stateful Ad Hoc routing:* Stateful ad hoc routing protocols require node to maintain some routing information that is collected using the routing protocol (e.g., through route request propagation or by reversing paths taken by the query). • *Stateless Geometric Ad Hoc routing:* These kinds of protocols only track the position of their neighbors and select among them a neighbor that is likely to be closer to the destination.
6.	Classification based on Epidemic behavior (Akdere et al., 2006)	• *Pull based epidemic algorithm:* A node asks a selected neighbor for new information. The node will receive new information only if the neighbor has new information. • *Push based epidemic algorithm:* A node with new information sends the information to a selected neighbor. • *Pull-push based epidemic algorithm:* This algorithm is a combination of two models described above.
7.	Classification based on Sensor Node Architecture (Al-Karaki & Kamal, 2004)	• *Protocols operating on flat topology (WSN consisting Homogeneous nodes)* • *Protocols operating on hierarchical topology(WSN consisting Heterogeneous nodes)*.
8.	Classification based on Protocol's initialization point (Biradar et al., 2009)	• *Source-initiated (Src-initiated):* A source-initiated protocol sets up the routing paths upon the demand of the source node, and starting from the source node. Here source advertises the data when available and initiates the data delivery. • *Destination-initiated (Dst-initiated):* A destination initiated protocol, on the other hand, initiates path setup from a destination node.
9.	Classification based on how the source finds the destination (Biradar et al., 2009)	• *Proactive:* A proactive protocol sets up routing paths and states before there is a demand for routing traffic. Paths are maintained even there is no traffic flow at that time. This approach is best suited for applications having fixed nodes • *Reactive:* In reactive routing protocol, routing actions are triggered when there is data to be sent and disseminated to other nodes. Here paths are setup on demand when queries are initiated. This approach is best suited for applications mobile nodes • *Hybrid:* This approach combines both techniques.
10.	Classification based on the basis of how to reduce useful energy consumption (Younis & Fahmy, 2004)	• Protocols that control the transmission power level at each node by increasing network capacity while keeping the network connected. • Protocols that make routing decisions based on power optimization goals. • Protocols that control the network topology by determining which nodes should participate in the network operation (be awake) and which should not (remain asleep).
11.	Cooperative routing (Castillo et al., 2007)	• In this approach, sensor nodes send data to a central node that join the data to reduce the cost in terms of energy consumption.

Table 15. (continues) Protocol classifications and sub-classifications for WSNs

to WSNs these design challenges are identified (Dwivedi et al., 2009a; Eriksson, 2009; Al-Karaki & Kamal, 2004; Karl & Willig, 2006; Akyildiz et al., 2002b; Akkaya & Younis, 2005; Wachs et al., 2007).

- Due to the relatively large number of sensor nodes, it is not possible to build a global addressing scheme for the deployment of a large number of sensor nodes as the overhead of ID maintenance is high. Thus, traditional IP based protocols may not be applied to WSNs.
- In contrast to typical communication networks, almost all applications of sensor networks require the flow of sensed data from multiple sources to a particular Base Station.
- Sensor nodes are tightly constrained in terms of energy, processing, and storage capacities. Thus, they require careful resource management.
- In most application scenarios, nodes in WSNs are generally stationary after deployment except for, may be, a few mobile nodes.
- Sensor networks are application specific, i.e., design requirements of a sensor network change with application.
- Position awareness of sensor nodes is important since data collection is normally based on the location.
- Finally, data collected by various sensors in WSNs is typically based on common phenomena; hence there is a high probability that this data has some redundancy.

Visibility (Wachs et al., 2007) is a new metric for WSNs protocol design. The objective of this visibility metric is that “Minimize the energy cost of diagnosing the cause of a failure or behavior”.

14. Major existing protocols for wireless sensor networks

A lot of protocols has been proposed in various research contributions, some of them are as follows: Rumor, DSR, SER (Stream Enabled Routing), AODV, SPIN (Sensor Protocols for Information via Negotiation) (SPIN-PP, SPIN-EC, SPIN-BC, SPIN-RL), GRAB, Direct Diffusion, GAF, SEER (Simple Energy-Efficient Routing), GBR, ARPEES, TIDD, TEEN, CADR, APTEEN, ACQUIRE, CEDAR, COUGAR, SAR, TinyAODV, PEQ (Periodic Event-driven and Query-based), GEAR, HPEQ (Hierarchical PEQ), MECN, CPEQ (Cluster PEQ), SMFCN, HEAP (Hierarchical Energy Aware Protocol for Routing & Aggregation in Sensor Networks), GF, PEGASIS (Power Efficient Gathering in Sensor Information System), GF-RSST, HPEGASIS (Hierarchical PEGASIS), LEACH, etc.

Some good research contributions (Al-Karaki & Kamal, 2004; Wachs et al., 2007) presents survey on existing WSN Protocols, whereas some other good one are dedicated to comparison, classification and other aspects of WSN Protocols (Dwivedi & Vyas, 2010; Biradar et al., 2009; Al-Karaki & Kamal, 2004; Wachs et al., 2007; Castillo et al., 2007).

15. Existing protocol classifications for wireless sensor networks

A careful attention is needed while selecting or proposing a new routing protocols for wireless sensor networks because WSNs are challenging due to the inherent characteristics such as energy efficiency and awareness, connection maintenance, minimum resource usage limitation, low latency, load balancing in terms of energy used by sensor nodes, etc. Various classifications for WSNs are presented in different literatures, at a glance these are (Table 15).

16. Protocol evaluation factors

These are the some parameters on which routing protocols must be evaluated during designing new one:

Evaluation Parameter	Description
Power Usage	Sensor node's lifetime is clearly dependent on its power source, thus useful power usage must be which involves: transmitting/receiving data, processing query requests, forwarding queries/data to neighboring nodes.
Data Aggregation	Substantial energy savings and traffic optimization can be obtained through data aggregation.
Scalability	The possibility to enlarge and reduce the network.
Reliability or Fault Tolerance	Fault tolerance is the ability to sustain WSN functionalities without any interruption due to node failures.
Latency (delay) and Overhead	Multi-hop relays and data aggregation cause data latency, these important factors influences routing protocol design.
Data Delivery Model	Data delivery model (Continuous, Event-driven, Query-driven , Hybrid) (Ahvar & Fathy, 2010) determines when the data collected by the sensor node has to be delivered.
Quality of Service (QoS)	Quality service required by the application, involves: length of life time, data reliability, energy efficiency, location-awareness, collaborative-processing, etc. QoS factors will affect the selection of routing protocols for a particular application.
Security	Security concerns needs special attention in current era where data stealing and data diddling becomes major issue.
Node Deployment option	Node deployment option affects the performance of routing protocol basically in terms of energy consumptions.
Topology	Topology of a WSN affects many of its characteristics like; latency, capacity, and robustness. As well as, the complexity of data routing and processing depends on the network topology.
Sensor Density and Network Size	Sensor density of nodes affects the degree of coverage area of interest whereas networks size affects reliability, accuracy, and data processing algorithms.
Environment or Scenario	A critical parameter, because node and network lifetime is directly dependent on it.
Byte Overhead (Saaranen & Pomalaza-Ráez, 2004)	Byte overhead means the total number of bytes in the routing control messages needed to find a route to the sink. For flooding, byte overhead means the total number of bytes in the extra messages flooded throughout the network. In both cases the bytes in the data packets transmitted by nodes along the route from the originating node to the sink node are not counted as overhead.

Table 16. WSN Protocol evaluation factors

Except these there are exist some common performance metrics, including latency, throughput, success rates, energy consumption and energy load, that must be calculated for the evaluation of routing algorithms.

17. Theoretical aspects of major energy efficient protocols

A classification on energy efficient/aware routing protocols is available in a research contribution (Ahvar & Fathy, 2010) which classified this type of protocols into: Energy Saver and Energy Manager. Energy saver protocols decrease energy consumption totally because most of them try to find the shortest path between source and destination to reduce energy consumption. The objective of energy manager protocols is to balance energy consumption in networks to avoid network partitioning. In first approach finding best route is totally based on energy balancing consideration, it may lead to long path with high delay and decreases network lifetime whereas in later approach finding best route only with the shortest distance consideration may lead to network partitioning. A lot of researches were conducted on the energy efficiency/awareness issue, some are presented here (Table 17)

18. Security issues in wireless sensor networks

In a survey paper (Dwivedi et al., 2009b) different classes of adversaries, and considers security goals in each scenario (indoor and outdoor) of WSNs, including: sensor nodes, networks of sensor nodes, operating systems, applications, middleware, and internet, are presented. This paper also presents valuable, in-depth recommendations of how to design and implement a security strategy for WSN. A procedure for protecting systems makes sure that the facility is physically secure, provides a recovery/restart capability, and has access to backup files establishing a priority sequence, one would probably want to start from within the firm and work out. Threats and their usual defenses are illustrated in (Figure 4)

Most WSN routing protocols are quite simple thus sometimes even more susceptible to attacks. Most network layer attacks against sensor networks falls under one of the following categories: Selective forwarding, Sinkhole attacks, Sybil attacks, Wormholes, HELLO flood attacks, Spoofed/Altered/Replayed routing information, Acknowledgement spoofing.

Some security issues that must need attention in wireless sensor networks, are as follows: Secure routing, Secure discovery and verification of location, Key establishment and trust setup, Attacks against sensor nodes, Secure group management, and Secure data aggregation.

In the ideal world, a secure routing protocol should guarantee the integrity, authenticity, and availability of messages in the presence of adversaries of arbitrary power. Every eligible receiver should receive all messages intended for it and be able to verify the integrity of every message as well as the identity of the sender. Several countermeasures and design considerations are also proposed in a research contribution (Karlof & Wagner, 2003).

Some mechanisms for authentication and security are based on public key cryptography. Public key cryptography is too expensive for sensor nodes. Security protocols for sensors networks must rely exclusively on efficient symmetric key cryptography. These protocols are too expensive in terms of node state and packet overhead and are designed to find and establish routes between any pair of nodes - a mode of communication not prevalent in sensor networks. Tackling with natural and manmade disasters is only possible with proper planning.

S.N.	Energy Efficient Protocol	Major Theoretical Aspects
1.	TEEN (Threshold sensitive Energy Efficient sensor Network protocol) (Manjeshwar & Agarwal, 2001)	- First protocol for reactive networks with enhanced efficiency. - Time critical data reaches the user almost instantaneously. Eminently well suited for time critical data sensing applications. - Message transmission consumes much more energy than data sensing. So, even though the nodes sense continuously, the energy consumption in this scheme can potentially be much less than in the proactive network, because data transmission is done less frequently. - The soft threshold can be varied, depending on the criticality of the sensed attribute and the target application. - A smaller value of the soft threshold gives a more accurate picture of the network, at the expense of increased energy consumption. Thus, the user can control the trade-off between energy efficiency and accuracy. - At every cluster change time, the attributes are broadcast afresh and so, the user can change them as required. - The main drawback of this scheme is that, if the thresholds are not reached, the nodes will never communicate; the user will not get any data from the network at all and will not come to know even if all the nodes die. Thus, this scheme is not well suited for applications where the user needs to get data on a regular basis. - Another possible problem with this scheme is that a practical implementation would have to ensure that there are no collisions in the cluster.
2.	APTEEN (Adaptive Periodic Threshold-sensitive Energy Efficient Sensor Network Protocol) (Manjeshwar & Agarwal, 2002)	- A Protocol for Hybrid network (inherit best characteristics of both proactive and reactive network). - To provide periodic data collection as well as near real-time warnings about critical events. - By sending periodic data, it gives the user a complete picture of the network. It also responds immediately to drastic changes, thus making it responsive to time critical situations. Thus, it combines both proactive and reactive policies. - It offers a flexibility of allowing the user to set the time interval (TC) and the threshold values for the attributes. - Energy consumption can be controlled by the count time and the threshold values. - The hybrid network can emulate a proactive network or a reactive network, by suitably setting the count time and the threshold values. - The main drawback of this scheme is the additional complexity required to implement the threshold functions and the count time. However, this is a reasonable trade-off and provides additional flexibility and versatility.
3.	HEED (Hybrid Energy-Efficient Distributed clustering) (Younis & Fahmy, 2004)	- An energy-efficient clustering protocol, using residual energy as primary parameter and network topology features (e.g. node degree, distances to neighbors) as secondary parameters. - Here all nodes are assumed to be homogenous nodes (with same initial energy). - It extends the basic scheme of LEACH. - The clustering process is divided into a number of iterations, as well as in each iteration nodes which are not covered by any cluster head doubles their probability of becoming a cluster head. - Since it enable every node to independently and probabilistically decide on its role in the clustered network, thus cannot guaranteed optimal elected set of cluster heads.

Table 17. Major theoretical aspects of some major energy efficient protocols for WSNs

S.N.	Energy Efficient Protocol	Major Theoretical Aspects
4.	H-HEED (Heterogeneous - HEED) (Kour & Sharma, 2010)	- A protocol for heterogeneous WSN. - Cluster head selection is primarily based on the residual energy of each node. Since the energy consumed per bit for sensing, processing, and communication is typically known, and hence residual energy can be estimated. - Intra cluster communication cost is considered as the secondary parameter to break the ties, tie means that a node might fall within the range of more than one cluster head. - Different level of heterogeneity is introduced: 2-level, 3-level and multi-level in terms of the node energy. - In 2-level H-HEED, two types of sensor nodes, i.e., the advanced nodes and normal nodes are used. - In 3-level H-HEED, three types of sensor nodes, i.e. the super nodes, advanced nodes and normal nodes are used. - In this heterogeneous approach all the sensor nodes are having different energy as a result nodes will die randomly. - Multi-level H-HEED prolongs lifetime and shows better performance than other level of H-HEED and HEED protocol.
5.	Reactive Energy Decision Routing Protocol (REDRP) (Ying-Hong et al., 2006)	- To solve the problem of limited energy, the loading of nodes have to be distributed as possible as it can. - If the energy consumption can be shared averagely by most nodes, then the lifetime of sensor networks will be enlarged. - This protocol will create the routes in reactive routing method to transmit the data node gathered. - It uses the residual energy of nodes as the routing decision for energy-aware.
6.	PEGASIS (Power-Efficient Gathering in Sensor Information Systems) (Lindsey & Raghavendra, 2002)	- A near optimal chain-based protocol and an enhanced descendant of LEACH. - It has two main objectives: increases the lifetime of each node by using collaborative techniques and allow only local coordination between nodes that are close together so that the bandwidth consumed in communication is reduced. - Nodes route data destined ultimately for the base station through intermediate nodes. - In determining the routes only consider the energy of the transmitter and neglect the energy dissipation of the receivers. - It assumes that each sensor node can be able to communicate with the base-station directly and all nodes maintain a complete database about the location of all other nodes in the network. - The method of which the node locations are obtained is not outlined. - It also assumes that all sensor nodes have the same level of energy and they are likely to die at the same time.
7.	Hierarchical-PEGASIS (Savvides et al., 2001)	- Its objective is to decrease the delay incurred for packets during transmission to the base-station. - In its concept only spatially separated nodes are allowed to transmit at the same time. - This chain-based protocol with CDMA capable nodes, constructs a chain of nodes, that forms a tree like hierarchy, and each selected node in a particular level transmits data to the node in the upper level of the hierarchy, that ensures data transmitting in parallel and reduces the delay significantly. - Results shows that this hierarchical extension of PEGASIS performs better than the regular PEGASIS scheme by a factor of about 60.
8.	SHPER (Scaling Hierarchical Power Efficient Routing) (Kandris et al., 2009)	- Enhanced integration of a hierarchical reactive routing protocol. - It supposes the coexistence of a base station and a set of homogeneous sensor nodes which are randomly distributed within a delimited area of interest. - Consists of two phases: the initialization phase and the steady state phase. - Hard and soft thresholds are utilized in the SHPER protocol as with TEEN. - Best suited in real life applications where imbalance in energy distribution is the common case. - Network scalability is retained because it adopts both multi-hop routing and hierarchical architecture.

Table 17. (continues) Major theoretical aspects of some major energy efficient protocols for WSNs

S.N.	Energy Efficient Protocol	Major Theoretical Aspects
9.	TREnD (Timely, Reliable, Energy-efficient and Dynamic) (Marco et al., 2010)	- A novel cross-layer WSN protocol for control applications. - The routing algorithm of TREnD is hierarchically subdivided into two parts: a static route at inter clusters level and a dynamical routing algorithm at node level. This is supported at the MAC layer by hybrid TDMA/CSMA solution. - The protocol parameters are adapted by an optimization problem, whose objective function is the network energy consumption, and the constraints are the reliability and latency of the packets. - It uses a simple algorithm that allows the network to meet the reliability and latency while minimizing for energy consumption. - It is best fit for industrial environments.
10.	LEACH (Low Energy Adaptive Clustering Hierarchy) (Heinzelman et al., 2000)	- A most popular cluster-based protocol, which includes distributed cluster formation. - The idea is to form clusters of the sensor nodes based on the received signal strength and use local cluster heads as routers to the sink. - It randomly selects a few sensor nodes as cluster-heads and rotates this role to evenly distribute the energy load among the sensors in the network. - Its operation is separated into two phases: setup phase where clusters are organized and CHs are selected and steady state phase where the actual data transfer to the base station takes place. - It uses a TDMA/CDMA MAC to reduce inter-cluster and intra-cluster collisions. - Optimal number of cluster heads is estimated to be 5% of the total number of nodes. - This protocol is most appropriate for the applications when there is a need for constant monitoring.
11.	SEER (Simple Energy Efficient Routing) (Hancke & Leuschner, 2007)	- A protocol that considers energy saving and balancing, with poor idea about energy balancing. - Once the network has been deployed in the area where it is to operate, the sink transmits a broadcast packet. - Each node in the network is assumed to have a unique address within the network. - When a node observes new data, it initiates the process of routing. Two types of data packets can be sent: normal data and critical data. - When nodes receive a data message they update the remaining energy value in the neighbor table for the neighbor that sent the message. Nodes that forward data messages follow the same process, except for minor differences. - If node's remaining energy falls below a certain threshold, it transmits an energy message to all of its neighbors to inform them of its energy level. - The sink node periodically sends a broadcast message through the network so that nodes can add new neighbors that joined the network to neighbor tables and remove neighbors that have failed from the neighbor tables. - Nodes also update remaining energy values stored in the neighbor tables.
12.	BEAR (Balanced Energy-Aware Routing) (Ahvar & Fathy, 2010)	- An extended version of SEER protocol with some visible difference specially in forwarding data procedure that saves and balance energy consumption in WSNs. - Finds optimal route in energy level and hop count both. - Routing decisions in BEAR are based on the distance to the base-station as well as on remaining battery energy level of nodes on the path towards the base station. - BEAR is better than the SEER protocol in energy managing, due to the fact that BEAR sends data packet along a balanced path.

Table 17. (continues) Major theoretical aspects of some major energy efficient protocols for WSNs

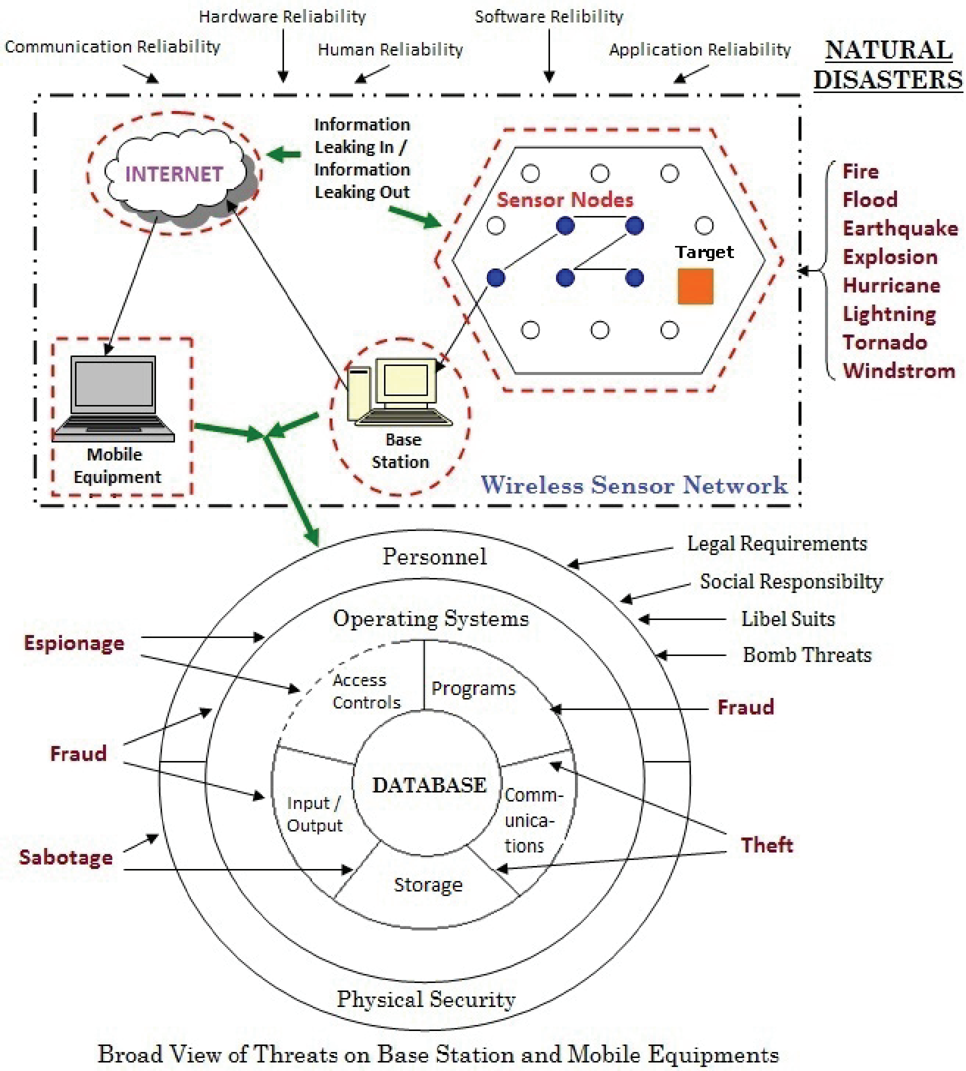


Fig. 4. Security threats and their usual defenses in Wireless Sensor Networks (Dwivedi et al., 2009b)

19. Reference

Dwivedi, A. K. & Vyas, O.P. (2010). Network Layer Protocols for Wireless Sensor Networks: Existing Classifications and Design Challenges, *International Journal of Computer Applications (IJCA)*, Vol.8, No.12, Article 6, pp. 30-34.

Tatiana, M. (2010). Mini Hardware Survey. Available from http://www.cse.unsw.edu.au/~sensar/hardware/hardware_survey.html

Imperial college London, U.K. (2007). Body Sensor Networks. Available from http://ubimon.doc.ic.ac.uk/bsn/index.php?m=206

Embedded WiSeNts Platform Survey (2006). Available from http://www.embedded-wisents.org/studies/survey_wp2.html

Senses (2005). Available from http://senses.cs.ucdavis.edu/resources.html

The Sensor Network Museum (2010). Available from http://www.snm.ethz.ch/Main/Homepage

Hempstead, M.; Lyons, M.J.; Brooks D. & Wei, G.-Y. (2008). Survey of Hardware Systems for Wireless Sensor Networks, *Journal of Low Power Electronics*, Vol.4, pp. 1-10.

Fröhlich, A.A. & Wanner L.F. (2008). Operating System Support for Wireless Sensor Networks, *Journal of Computer Science*, Vol.4, No.4, pp. 272-281.

Reddy, A.M.; Kumar, V.A.V.U.P.; Janakiram, D, & Kumar, G.A. (2007). Operating Systems for Wireless Sensor Networks: A Survey, *Technical Report*, IIT Madras, Chennai, India.

Dwivedi, A. K.; Tiwari, M.K. & Vyas, O.P. (2009). Operating Systems for Tiny Networked Sensors: A Survey, *International Journal of Recent Trends in Engineering*, Vol.1, No.2, pp. 152-157.

Manjunath, D. (2007). A Review of Current Operating systems for Wireless Sensor Networks, *Technical Report*, Department of ECE, Indian Institute of Science, Bangalore, INDIA.

Akyildiz, I.; Su, W.; Sankarasubramaniam, Y. & Cayirci, E. (2002). A survey on Sensor Networks, *IEEE Communications Magazine*, Vol.40, Issue:8, pp. 102-114.

Karl, H. & Willig, A. (2003). A Short Survey of Wireless Sensor Networks, *TKN Technical Report TKN-03-018*, Technical University Berlin, Germany.

Dulman, S. & Havinga, P. (2005). Architectures for Wireless Sensor Networks, *Proceedings of the IEEE ISSNIP 2005*, pp. 31-38.

NeTS-NOSS: Creating an Architecture for Wireless Sensor Networks, (2007). Available from http://snap.cs.berkeley.edu/documents/architecture.pdf

Vazquez, J.; Almeida, A.; Doamo, I.; Laiseca, X. & Orduña, P. (2009). Flexeo: An Architecture for Integrating Wireless Sensor Networks into the Internet of Things, *Proceedings of 3rd Symposium of Ubiquitous Computing and Ambient Intelligence 2008*, Springer Berlin/Heidelberg, Vol.51, pp. 219-228, 2009.

Schott, W.; Gluhak, A.; Presser, M.; Hunkeler U. & Tafazolli, R. (2007). e-SENSE Protocol Stack Architecture for Wireless Sensor Networks. *Proceedings of 16th IST Mobile and Wireless Communication Summit*, pp. 1-5.

Biradar, R.V.; Patil, V.C.; Sawant, S.R. & Mudholkar, R.R. (2009). Classification and Comparison of Routing Protocols in Wireless Sensor Networks, *Special Issue on Ubiquitous Computing Security Systems, UbiCC Journal*, Vol.4, pp. 704-711.

Katiyar, V.; Chand, N. & Chauhan, N. (2010). Recent advances and future trends in Wireless Sensor Networks, *International Journal of Applied Engineering Research*, Vol.1, No.3, pp. 330-342, ISSN 0976-4259.

Distributed Computing Group's Sinalgo-Simulator for Network Algorithms. (2009). Available from http://disco.ethz.ch/projects/sinalgo/tutorial/tuti.html

Eriksson, J. (2009). Detailed Simulation of Heterogeneous Wireless Sensor Networks, *Dissertation for Licentiate of Philosophy in Computer Science*, Uppsala University, Sweden, ISSN 1404-5117.

Shu, L.; Hauswirth, M.; Zhang, Y.; Mao, S.; Xiong N. & Chen, J. (2009). NetTopo: A Framework of Simulation and Visualization for Wireless Sensor Networks, *Proceedings of the ACM/Springer Mobile Networks and Applications*.

Parbat, B.; Dwivedi A.K. & Vyas, O.P. (2010). Data Visualization Tools for WSNs: A Glimpse, *International Journal of Computer Applications*, Vol.2, No.1, pp.14-20, ISSN 0975-8887.

Kosucu, B.; Irgan, K.; Kucuk, G.; & Baydere, S. (2009). FireSenseTB: A Wireless Sensor Networks Testbed for Forest Fire Detection, *Proceedings of the* IWCMC.

Tavakoli, A. (2007). Wringer: A Debugging and Monitoring Framework for Wireless Sensor Networks, *Proceedings of the ACM SenSys Doctoral Colloquium*.

Panta, R.K.; Bagchi, S. & Midkiff, S.P. (2009). Zephyr: Efficient Incremental Reprogramming of Sensor Nodes using Function Call Indirections and Difference Computation, *Proceedings of the USENIX*. Available from http://www.usenix.org/events/usenix09/tech/full_papers/panta/panta_html

Dwivedi, A.K.; Patle, V.K. & Vyas, O.P. (2010) Investigation on Effectiveness of Simulation Results for Wireless Sensor Networks, *CCIS*, Springer-Verlag Berlin Heidelberg, Vol.70, pp. 202-208.

Li, Q.; Österlind, F.; Voigt, T.; Fischer, S. & Pfisterer, D. (2010). Making Wireless Sensor Network Simulators Cooperate, *Proceedings of the PE-WASUN'10*, Bodrum, Turkey, pp. 95-98.

Al-Karaki, J.N. and Kamal, A.E. (2004). Routing Techniques in Wireles Sensor Networks: A Survey, *IEEE Wireless Communications (J)*, pp. 06-28.

Karl, H. and Willig, A. (2006). Protocols and Architectures for Wireless Sensor Networks, *Editorial John Wiley & Sons Ltd.*, ISBN 13978-0-470-09510-2.

Akyildiz, I.F.; Su, W.; Sankarasubramaniam, Y. & Cayirci, E. (2002). Wireless Sensor Networks: A Survey, *Computer Networks (Elsevier) (J)*, Vol.38, pp.393-422.

Akkaya, K. and Younis M. (2005). A Survey on Routing Protocols for Wireless Sensor Networks, *Ad Hoc Network (Elsevier) (J)*, Vol.3, pp. 325-349.

Wachs, M.; Choi; Jung, J. II; Lee, W.; Srinivasan, K.; Chen, Z.; Jain, M. and Levis, P. (2007). Visibility: A New Metric for Protocol Design. *Proceedings of ACM SenSys*.

Olivares, T.; Tirado, P.J.; Royo, F.; Castillo, J.C. & Orozoco-Barbosa, L.: (2007). IntellBuilding: A Wireless Sensor Network for Intelligent Buldings. Poster. *Proceedings of 4th European Conference on Wireless Sensor networks (EWSN)*.

Akdere, M.; Bilgin, C.C.; Gerdaneri, O.; Korpeoglu, I.; Ulusoy, Ö. and Cetintemel, U. (2006). A Comparison of Epidemic Algorithms in Wireless Sensor Networks. *Elsevier Journal of Computer Communications*, Vol.29, pp. 2450-2457.

Younis, O. & Fahmy, S. (2004). HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks. *IEEE transactions on Mobile Computing*, Vol.3, pp. 366-379.

Castillo, J.C.; Olivares, T. & Orozco-Barbosa, L. (2007). Routing Protocols for Wireless Sensor Networks-Based Network. *Technical Report*, Albacete Research Institute of Informatics, University of Castilla, SPAIN.

Tilak, S.; Abu-Ghazaleh, N.B. & Heinzelman, W. (2002). A Taxonomy of Wireless Microsensor Network Models, *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol.6, Issue 2, pp. 28–36.

Saaranen, A. & Pomalaza-Ráez, C.A. (2004). Comparison of Reactive Routing and Flooding in Wireless Sensor Networks, *Proceedings of Nordic Radio Symposium*, Oulu, Finland.

Ahvar, E. & Fathy, M. (2010). BEAR: A Balanced Energy-Aware Routing Protocol for Wireless Sensor Networks. *Wireless Sensor Network*, Vol.2, pp. 793-800.

Manjeshwar, A. & Agarwal, D.P. (2001). TEEN: a Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks. *Proceedings of 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing*.

Manjeshwar, A. & Agarwal, D.P. (2002). APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless Sensor Networks. *Proceedings of International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 195-202.

Kour, H. & Sharma, A.K. (2010). Hybrid Energy Efficient Distributed Protocol for Heterogeneous Wireless Sensor Network. *International Journal of Computer Applications*, Vol.4, pp. 1-5.

Ying-Hong, W.; Yi-Chien, L.; Ping-Fang, F. & Chih-Hsiao, T. (2006). REDRP: Reactive Energy Decisive Routing Protocol for Wireless Sensor Networks. *Ubiquitous Intelligence and Computing, LNCS*, Vol.4159, pp. 527-535, Springer Berlin/Heidelberg.

Lindsey, S. & Raghavendra, C. (2002). PEGASIS: Power-Efficient Gathering in Sensor Information Systems, *Proceedings of IEEE Aerospace Conference*, Vol.3, pp. 1125-1130.

Savvides, A.; Han, C-C & Srivastava, M. (2001). Dynamic Fine-grained localization in Ad-Hoc Networks of Sensors. *Proceedings of 7th ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 166-179.

Kandris, D.; Tsioumas, P.; Tzes, A.; Nikolakopoulos, G. & Vergados, D.D. (2009). Power Conservation through Energy Efficient Routing in Wireless Sensor Networks. *Sensors*, 9, pp. 7320-7342, ISSN 1424-8220.

Marco, P.D.; Park, P.; Fischione, C. & Johansson, K.H. (2010). TREnD: a Timely, Reliable, Energy-efficient and Dynamic WSN Protocol for Control Applications. *Proceedings of Information Communication Conference*.

Heinzelman, W.; Chandrakasan, A. & Balakrishnan, H. (2000). Energy-Efficient Communication Protocol for Wireless Microsensor Networks. *Proceedings of 33rd Hawaii International Conference on System Sciences (HICSS '00)*.

Hancke, G.P. & Leuschner, C.J. (2007). SEER: A Simple Energy Efficient Routing Protocol for Wireless Sensor Networks, *South African Computer Journal*, Vol.39, pp.17-24.

Dwivedi, A.K.; Tiwari, M.K. & Vyas, O.P. (2009). A Review of Security in Wireless Sensor networks for Indoor Application Scenario: Prospects and Challenges, *Proceedings of National Conference on Wireless Communication and Networking (WINCON)*, pp. 138-148.

Karlof, C. & Wagner, D. (2003). Secure routing in wireless sensor networks: Attacks and countermeasures. *Proceedings of 1st IEEE International Workshop on Sensor Network Protocols and Applications*.

Dwivedi, A.K. & Vyas, O.P. (2011). An Exploratory Study of Experimental Tools for Wireless Sensor Networks. *Wireless Sensor Network*,Vol. 3, ISSN 1945-3078 (Print), 1945-3086 (Online). Available from http://www.scirp.org/journal/wsn

15

Software Defined Radio Platform for Cognitive Radio: Design and Hierarchical Management

Amor Nafkha, Christophe Moy, Pierre Leray,
Renaud Seguier and Jacques Palicot
SUPELEC/IETR, Avenue de la Boulaie,
Cesson Sévigné Cedex,
France

1. Introduction

Cognitive Radio (CR) Mitola (2000) is a promising technology to improve spectrum utilization of wireless communication systems. Current investigations in CR have been focused on the physical layer functionality. The cognitive radio, built on a software-defined radio, assumes that there is an underlying system hardware and software infrastructure that is capable of supporting the flexibility needed by the cognitive algorithms. As already foreseen by Mitola Mitola & Maguire (1999), a Cognitive Radio is the final point of Software Defined Radio (SDR) platform evolution: *a fully reconfigurable radio that changes its communication functions depending on network and/or user demands*. Mitola's definition on reconfigurability is very broad and we only focus here on the reconfigurability of the hardware platform for Cognitive Radio. SDR basically refers to a set of techniques that permit the reconfiguration of a communication system without the need to change any hardware system element. As explained in the schematic of figure 1, this relies on a cognitive circle. Figure 1 (a) is from Mitola (2000) and figure 1 (b) is a simplified view of the cycle summarized in three main steps:

- Observe: gathers all the sensing means of a CR,
- Decide: represents all that implies some intelligence including learning, planning decision taking,
- Adapt: reconfigures the radio, designed with SDR principles, in order to be as flexible as possible.

The figure 2 draw the general approach that can help the radio to better adapt its functionality for a given service in a given environment without restriction on the sensors nature.
Sensors are classified in function of the OSI layers they correspond to, with a rough division in three layers. Corresponding to the lower layers of the OSI model, we find specifically all the sensing information related to the physical layer: propagation, power consumption, coding scheme, etc. At the intermediate level are all the information that participate to vertical handover, or can help to make a standard choice, as a standard detection sensor for instance. The network load of the standards supported by the equipment may also be of interest. It also includes the policies concerning the vicinity, the town or the country. The highest layer is related to the applications and all that concerns the human interaction

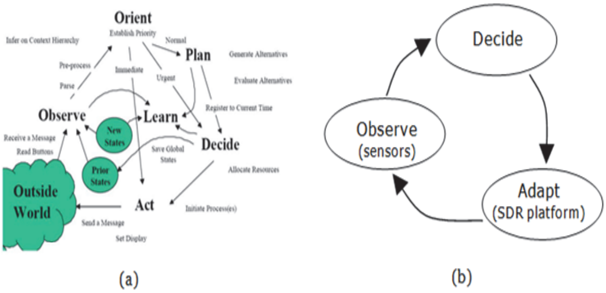
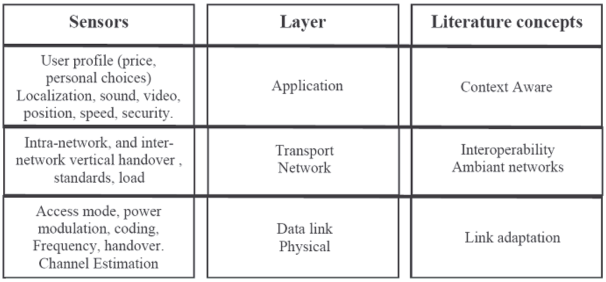


Fig. 1. (a) Mitola’s cognitive cycle, (b) simplified version



Sensors	Layer	Literature concepts
User profile (price, personal choices) Localization, sound, video, position, speed, security.	Application	Context Aware
Intra-network, and inter-network vertical handover , standards, load	Transport Network	Interoperability Ambiant networks
Access mode, power modulation, coding, Frequency, handover. Channel Estimation	Data link Physical	Link adaptation

Fig. 2. Simplified OSI model for cognitive radio context

with the communicating device. It is related to everything that concerns the user, his habits, preferences, policies, profile. If a user has the habit to connect to a video on demand service every evening while coming back home from office by metro, a CR terminal should be aware of it to plan all the requirements in terms of battery life, sufficient quantity of credit on his contract, vertical handover succession depending on each area during the trip, etc. The equipment can be aware of its environment with the help of sensors like microphone, video-camera, bio-sensors, etc. As we are at the early beginning of such technology, it is difficult to foresee all the possibilities. We can think, for instance, that user’s biometric information and/or facial recognition will ensure equipment security. Video-camera could also be used to indicate if the terminal is outside or inside a building. This may impact propagation features, but also the capability or not to receive GPS signals. Another example could be given in the context of video conferencing, a separation between the face of the speaker and the background could help decreasing the data rate while refreshing slowly the background of the image Nafkha et al. (2007).

Note that this classification is also related to three well-known concepts of the literature:

- Context aware for higher layers Chen & Kotz (2000),
- Interoperability for intermediate layers Aarts et al. (2001),
- Link adaptation for lower layers Qiu & Chuang (1999).

All this may be combined to achieve cross-layer optimizations. This is one of the responsibilities of the cognitive engine in our mind. However, due to the high financial pressure on spectrum issues, CR is often restricted in the research community to spectrum management aspects as in Fan et al. (2008); Ghozzi et al. (2006). Opportunistic spectrum

access approaches are explored to increase the global use of the spectrum resources. FCC has been already opening the door for several years, in the TV broadcasting bands, and permits secondary users (e.g. not licensed) to occupy primary users spectrum when available.
More futuristic CR scenarios may also be considered concerning the spectrum management. We may even imagine in the very long term a fully deregulated spectrum access where all radio connections features would be defined on-the-fly: carrier frequency, modulation, data rate, coding scheme, etc. But this means also to overcome regulatory issues in addition to technological challenges.

2. Background and related work

The objective behind this section is to highlight other cognitive radio platforms and to give our architecture purpose.

2.1 Related work

There are a large number of experimental SDR platforms that have been developed to support individual research projects. The various experimental SDR platforms have made different choices in how they are addressed the issues of flexibility, partitioning and application. To highlight the variety of architectures, five popular platforms will be discussed briefly prior to introducing our platform.

- *NICT SDR Platform*: The Japanese National Institute of Information and Communications Technology (NICT) constructed a software defined radio platform to trial next generation mobile networks. The platform has two embedded processors, four Xilinx Virtex2 FPGA and RF modules that could support *1.9* to *2.4* and *5.0* to *5.3* GHz. The signal processing was partitioned between the CPU and the FPGA, with the CPU taking responsibility for the higher layers. An objective of this platform was to explore selection algorithms to manage handover between existing standards. To this end, a number of commercial standards were implemented, for example *802.11a/b/g*, digital terrestrial broadcasting, *WCDMA* and a general *OFDM* communication scheme.
- *Berkeley Cognitive Radio Platform*: This platform is based around the Berkeley Emulation Engine (BEE2) which is a platform that contains five high-powered Xilinx Virtex2 FPGAs and can connect up to eighteen daughter-boards. In the Cognitive Radio Platform radio, daughter-boards have been designed to support up to *25* MHz of bandwidth in an *85* MHz range in the *2.4* GHZ ISM Band. The RF modules have highly sensitive receivers and to avoid self-generated noise operate either concurrently at different frequencies (FDD) or at the same frequency in a time-division manner. This cognitive radio platform requires only a low-bandwidth connection to a supporting PC as all signal processing is performed on the platform.
- *Kansas University Agile Radio (KUAR)*: The KUAR platform was designed to be a low-cost experimental platform targeted at the frequency range *5.25* to *5.85* GHz and a tunable bandwidth of *30* MHz. The platform includes an embedded *1.4* GHz general purpose processor, Xilinx Virtex2 FPGA and supports gigabit Ethernet and PCI-express connections back to a host computer. This allows for all, or almost all processing to be implemented on the platform.

- *OpenAirInterface*: The mobile communications department at EURECOM proposed an open-source hardware/software development platform and open-forum for innovation in the area of digital radio communications. OpenAirInterface implements in software the Physical and Medium-access layers for wireless communications as well as providing a IPv4/IPv6/MPLS network device interface under Linux. The initiative targets 4^{th} generation wireless systems (UMTS Longterm-evolution (LTE), 802.16e/j) and rapidly-deployable MESH networks using a similar radio interface technologies. The development can be seen as an open-source testbed for advanced algorithmic prototyping and performance evaluation.
- *Universal Software Radio Peripheral (USRP)*: The USRP is one of the most popular SDR platforms currently available and it provides the hardware platform for the GNU Radio project. The first USRP system, released in *2004*, was a USB connected to a computer with a low-performance FPGA. The FPGA was used primarily for routing information but also allowed some limited signal processing. The USRP could realistically support about *3* MHz of bandwidth due primarily to the performance restrictions of the USB interface. The second generation platform was released in September *2008* and utilizes gigabyte Ethernet to allow support for *25* MHz of bandwidth. The system includes a medium range Xilinx Spartan3 device which allows for a local processing. The radio-frequency performance of the USRP was limited and is more directed towards experimentation rather than matching any communications standard.

Our proposed platform has been developed in order to achieve high flexibility and reconfigurability of the wireless baseband processing. For the hardware part, for example, we exploited the ability to reconfigure partial areas of an FPGA anytime after its initial configuration. Our development concerns all processing blocs: from the video treatment to the intermediary frequency signal generation. Our intention is not to develop any commercial platform, but just to test and verify our approach to achieve baseband flexibility using:

- Partial Reconfiguration Nafkha et al. (2007) and Common Operator Alaus et al. (2008).
- Hiearchical Reconfiguration Management Delahaye et al. (2005).
- Hierarchical and Distributed Cognitive Radio Architecture Management Godard. (2009).

2.2 The proposed solution

The proposed solution is a design approach and not a hardware platform itself so that it is not restricted to a specific hardware platform. It intends to answer the design issue of SDR in the following context:

- flexible processing including partial FPGA reconfiguration and Common Operator approach.
- heterogeneous processing, including processors (GPP, DSP), FPGA and ASICs,
- portability from a HW device target to another.

In order to cope with these characteristics, a modular-based approach is privileged. This is the main support of flexibility. It permits indeed to separate the radio application into sub-pieces that can be split in any sub-set depending on the HW devices that compute their processing needs. This also favorites changes in the repartition of the processing modules on the HW devices. As all processing modules are designed independently in a modular-based approach, this also guarantees the non dependence of processing modules in terms of operating rhythm. One can just not make them run faster than their fastest speed, but anything lower is compatible.

This is very straightforward in a processor environment as the processing modules varies with the processor frequency (or it architecture after compilation). But this is generalized to the reconfigurable HW world while using Globally Asynchronous Locally Synchronous (GALS) principles. It turns HW processing as SW in the sense that the exchanges between processing modules are asynchronous from the data rate they have to process. The consequence is that these processing modules can be ported to several designs at different speeds, with no dependence with the speed of other blocks. Another major effect is that it becomes transparent to replace a SW processing module, e.g. running in a processor, by a HW processing module, e.g. running in a FPGA, and vice versa. Moreover, a HW processing module can be easily moved to a processor instantiated inside a FPGA (such as a NIOS for Altera or a MicroBlaze for Xilinx) without reconsidering the global behavior of the processing modules it is connected to.

This design approach is completely compatible with an Intellectual property (IP) oriented design strategy. Re-usability has several major advantages: gains of time at all development stage, debug and validation stages, and integration stage. It permits also to benefit from third party expertise to speed-up or complete the proprietary designs. To sum-up the proposed solution consists in declaring rules for the design of IPs or processing modules so that they can be easily assembled in the design framework that is detailed below.

3. System structure

The presented real-time platform provides a simple wireless video stream broadcasting system to verify and test our approach. It consists of one transmitter as the base station and one receiver as the terminal. The system architecture is depicted in figure 3. Basically, the transmitter and receiver hosts can communicate and exchange their data through an existing TCP/IP networks or Intermediate Frequency (IF) link. The transmitter host utilizes USB port to communicate with the video camera. At the receiver side, any standard display monitor allows us to display the incoming video stream.

3.1 Hardware architecture

The digital hardware setup of the whole system is based on Sundance modules. The transmitter and receiver side contain a Sundance SMT310Q carrier boards, plugged via Peripheral Component Interconnect (PCI) bus to a standard PC. The hardware architecture is depicted in figure 4.

At the transmitter side shown at figure 5, the Sundance SMT310Q carrier-board is used to carry the processing modules (SMT395, SMT348 and SMT350 ADC/DAC) in the four available TIM-40 slots. The Sundance SMT395 module is placed in the first TIM-slot and controls the operation of other modules. It consists principally a Texas Instruments (TI) 6416T fixed-point Digital Signal Processor (DSP) running at 1 GHz, a Xilinx Virtex II Pro FPGA, and two Sundance High-speed Bus (SHB, up to 400MB/s) for fast data exchange with the other modules. In our platform the DSP is used as a control device for the ADC/DAC and memory modules and to set the parameters for the pre-distortion filter running in real-time on the FPGA at the module SMT350. Based on the Xilinx Virtex4 range, the SMT348 features 16MB of blistering fast QDRII memory, ensuring ample capacity to develop todays demanding applications. The SMT348 includes SHB and SLB (Sundance LVDS Bus) interfaces. It provides quick and easy connection to rapid ADC and DAC modules for data acquisition or software radio systems. The SMT350 module, is composed of:

- Two DACs DAC5686 from Texas Instrument with 16 bits of resolution and a maximum sampling rate of 500MSPS with interpolation filters
- Two ADCs ADS5500 from Texas Instrument with 14bits of resolution and a maximum sampling frequency of 125MHz
- A CDCM7005 from Texas Instrument which provides individual sample frequency to each converters

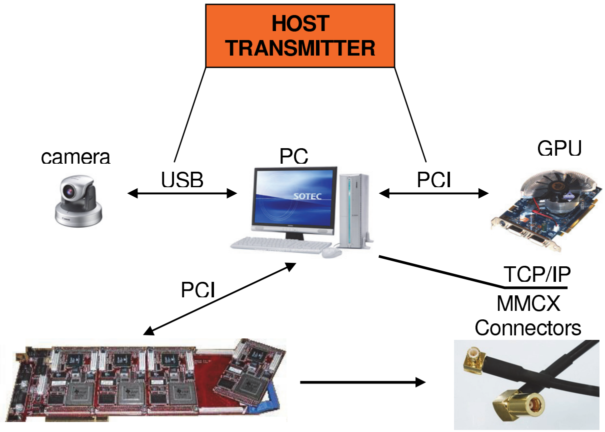


Fig. 3. Hardware Architecture

The stream server program encapsulates the video data into Internet Protocol (IP) stream and saves the IP stream in the buffer allocated in the main memory of the host PC. The DSP module fetches the data in the buffer through the PCI interface provided by the above mentioned carrier board and then executes the partial part of the digital baseband and Intermediate Frequency (IF) signal processing algorithms of the transmitter. The driver of the carrier board offers the DSP module the methods to access the main memory of the host PC through PCI interface by providing C/C++ Application Program Interface (API) functions. The Xilinx FPGA on the DSP module takes care of the Sundance High speed Bus (SHB) interfacing between the DSP module SMT395 and the FPGA module SMT348. The SHB interface is able to transfer 32-bit data at a 100 MHz clock. Via SHB the digital IF signal is forwarded to the SMT350 to generate the analog signal using its integrated Digital to Analog Converter (DAC). The analog IF signal goes through the low-pass filter.
The hardware setup of the receiver is similar to the transmitter, as shown in figure 5. In this case, the SMT350 is configured as an Analog to Digital Converter (ADC) module and the signal experiences the reciprocal of the transmitter. The IF signal coming from the transmitter is sampled synchronously by the ADC on the SMT350 module. The FPGA module SMT348 receives the digital samples from the SHB interface and accomplishes a high parallel part of

digital signal process algorithms of the receiver. The simplest part of the baseband process is sent to the SMT395 module via the SHB.

The final received IP packets are saved to the buffer in the main memory of the host PC through the PCI interface. The network layer program fetches the IP packets from the buffer and emits them to the certain IP port by IP socket programming. The video stream player always listens to the IP port and plays the video back.

In both side, transmitter/receiver, the testbed platform contains a Graphics Processor Unit (GPU). The main reason behind is that the GPU is specialized in compute intensive, highly parallel computation and therefore is designed in such a way that more transistors are devoted to data processing rather than data caching and flow control.

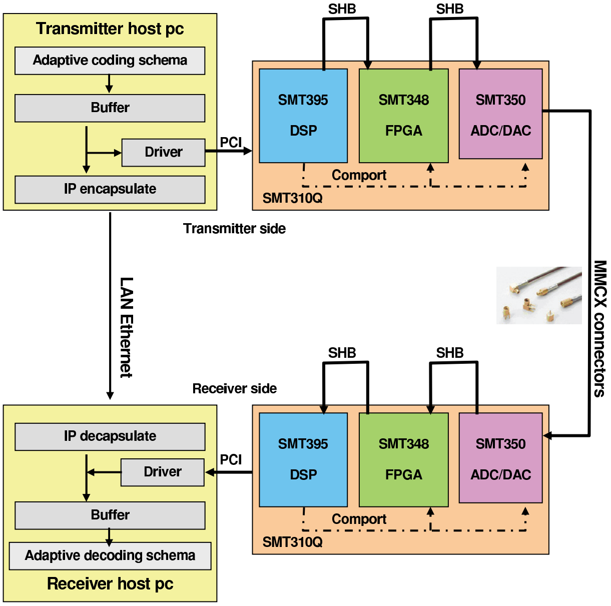


Fig. 4. Transmitter and receiver overview

3.2 Software architecture

The two host stations are a standard personal computers running Microsoft Windows XP and Microsoft Visual C++. Several hardware and software tools, as depicted in figure 6, are necessary for the completion of our testbed. These tools include the physical DSP/FPGA and their associated development board that allowed for continual reprogramming of test systems as well as many features for data storage and output display. Xilinx has also supplied a suite of tools that are used in our platform. These Xilinx software tools are used for developing the hardware and software aspects of the system. Although many of these tools have included documentation from Xilinx, their support of partially reconfigurable systems is currently somewhat lacking. Therefore, the integration of these tools into a working tool flow to achieve the goal of a partial reconfiguration for example required research from numerous sources and some experimentation with the tools. For the partial reconfiguration implementation, we need the following Xilinx tools:

- *Xilinx EDK* provides a framework for design of hardware/software components of the embedded processor systems on programmable logic. Appropriate tools for each stage of the design in addition to IBM PowerPC and Xilinx MicroBlaze processor cores infrastructure and peripheral IP cores facilitate hardware/software partitioning and design reuse.
- *Xilinx ISE* allows for complete FPGA development. It can automatically interpret the HDL syntax, synthesize the description, place and route the logic elements and then provide a software BIT file description to connect these logic elements together to create the circuit described in the HDL. All these tool flow steps require their own respective application program to perform the function.
- *Xilinx PlanAhead* is a floor-planning tool provided by Xilinx to allow developers flexibility on how there synthesis designs should be placed on the FPGA floorplan. This tool is useful in ASIC designs where locations of logic elements are an important factor to the performance of the application. For the scope of our platform, PlanAhead has partial reconfiguration options which make it a required tool in the partial reconfiguration tool flow.

The programming of the TI 6416T fixed-point DSP is achieved using the Code Composer Studio (CCS) Integrated Development Environment (IDE). The CCS IDE allows a user to connect, program in C, and run the DSP through a graphical user interface. furthermore, a user is able to view the memory contents of the DSP and profile the execution time for pieces of their code all in real-time. For high level synthesis of the digital signal process algorithms, we use the SynDEx tool environment Grandpierre et al. (1999) which provides a formal framework based on graphs and system-level computer-aided design (CAD) software. On the one hand, this tool specifies the functions of the applications, the distributed resources in terms of processors and/or specific integrated circuits, and communication media. On the other hand, it assists the designer in implementing the functions onto the resources while satisfying timing requirements and, as far as possible, minimizing the resources. The results is a real-time behavior of the application functions executed on various resources, like processors, integrated circuits or communication media. For the software part of the application, code is automatically generated as a dedicated real-time executive.

As a result of demand for video treatment at the both side (transmitter/receiver), the CUDA programming model is used. It is ANSI C extended by several keywords and constructs. The GPU is treated as a co-processor that executes data-parallel kernel code. The user supplies a single source program encompassing both host (CPU) and kernel (GPU) code. Each CUDA program consists of multiple phases that are executed on either the CPU or the GPU. The phases that exhibit little or no data parallelism are implemented in host (CPU), which is expressed in ANSI C and compiled with the host C compiler. The phases that exhibit rich data parallelism are implemented as kernel functions in the device (GPU) code. A kernel function defines the code to be executed by each of the massive number of threads to be invoked for a data-parallel phase. These kernel functions are compiled by the NVIDIA CUDA C compiler and the kernel GPU object code generator. There are several restrictions on kernel functions: there must be no recursion, no static variable declarations, and a non-variable number of arguments. The host code transfers data to and from the GPU's global memory using API calls. Kernel code is initiated by performing a function call.

SynDEx tools Grandpierre et al. (1999) provides a formal framework based on graphs and system-level software. On the one hand, these specify the functions of the applications, the distributed resources in terms of processors and/or specific integrated circuit and communication media, and the non-functional requirements such as real-time performances. On the other hand, they assist the designer in implementing the functions onto the resources while satisfying timing requirements and, as far as possible, minimizing the resources. This is achieved through a graphical environment (see figure 5), which allows the designer to explore manually and/or automatically the design space solutions using optimization heuristics. Exploration is mainly carried out through timing analysis and simulations. The results of these prediction's is a real-time behavior of the application functions executed on various resources, like processors, integrated circuits or communication media. This approach conforms to the typical hardware/software co-design process. Finally, for the software part of the application, code is automatically generated as a dedicated real-time executive.

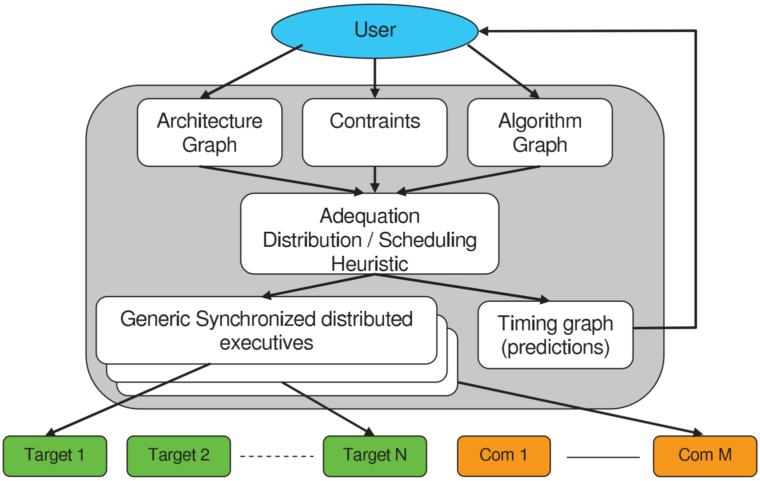

User
Architecture Graph
Contraints
Algorithm Graph
Adequation Distribution / Scheduling Heuristic
Generic Synchronized distributed executives
Timing graph (predictions)
Target 1
Target 2
Target N
Com 1
Com M

Fig. 5. SynDEx utilization global view

4. Hardware development

Even though the prototyping effort is focused on an FPGA-based design, we are also exploring the architectural benefits of custom integrated circuitry, primarily related to power consumption and the silicon area, which are important performance parameters for hardware designs used in mobile/portable platforms. The approach we have chosen to take involves identifying the hardware architecture appropriate for low-power configurable design based on heterogeneous blocks (i.e. blocks that are highly optimized for a particular function, yet flexible enough to support a variety of configuration parameters) as a compromise for the trade-off between programmability and power consumption/area. In addition to fast prototyping, the additional benefits of using modern FPGAs (e.g. Xilinx Virtex 4) are the availability of highly optimized features implemented as non-standard configurable logic blocks (CLB) like phase-locked loops, low-voltage differential signal, clock data recovery, lots of internal routing resources, hardware multipliers for DSP functions, memory, programmable

I/O, and microprocessor cores. These advantages simplify mapping from hierarchical blocks to FPGA resources.

4.1 Common operator

At the foundation of our study is the ′common operators′ technique to the design of reconfigurable equipment. Its main principle is the identification and (re)use throughout the design of common components that can each match several processing contexts, via a simple parameter adjustment. This technique can greatly increase the efficiency of a multi-standard software-defined radio, both in terms of its cost, and of the speed of reconfiguration during operation. The common operator technique belongs to the parameterization techniques firstly proposed by Jondral et al. (2002). The common operators′ technique is discussed more extensively in Alaus et al. (2008).

A part of the theoretical approach of this technique is presented in Rodriguez et al. (2007). Two different parameterization techniques have been proposed in the literature:

- The Common Function (CF) Technique consists in seeking an optimized generic function (the expected common function) like coding, mapping, among others which can replace the initial task present in a predefined set of standards. This Common Function (CF) technique was historically the first one proposed by several articles from Karlsruhe University (Germany) Jondral et al. (2002), Rhiemeier (2002)
- The Common Operator (CO) technique claims to be independent of the standards by finding the smallest set of highest-level operators like MAC, FFT, etc., which are used by the maximum functions number. It is an open technique. The foundation paper was Palicot & Roland (2003) which identified the FFT as a common operator

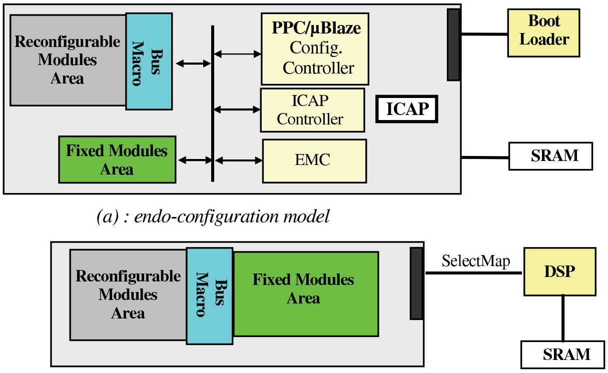
The CO technique is implemented in our platform in order to optimize both the area and the reconfiguration time on the FPGA.These operators are very small regards to the needed bit-stream, therefore the time to reconfigure a function using an operator is very small too. These operators are managed by the lowest level of our hierarchical reconfiguration manager(see section 5). This manager level is very close to the operator itself and in most cases embedded in the same resource. Furthermore, thanks to the partial reconfiguration approach, it is very easy to modify either a complete specific operator or parameters of an operator, providing a huge gain in reconfiguration time.

4.2 Partial Dynamic Reconfiguration

Partial Dynamic Reconfiguration is the capability to reconfigure specific areas of the FPGA at run-time after its initial configuration. PDR is carried out to allow the FPGA to adapt to changing hardware algorithms, improve resource utilization, to enhance performance or to reduce power consumption. In March of 2006, Xilinx introduced the early access partial reconfiguration flow along with the introduction of slice based bus macros which are pre-routed intellectual property (IP) cores. The restriction of full column modular PDR was removed allowing reconfigurable modules of any arbitrary rectangular size to be created. The EAPR flow also allows signals from the static regions to cross through the partially reconfigurable regions via the bus macros. Using the principle of glitch-less reconfiguration, no glitches will occur in signal routes as long as they are implemented identically in every reconfigurable module for a region. The only limitation of this approach is that all the partial bit-streams for a module, to be executed on a reconfigurable region, must be predetermined.

The Virtex-II and the Virtex-II Pro are the first Xilinx architectures that support Internal configuration access port (ICAP) Blodget et al. (2003) which is a subset of the Xilinx SelectMAP interface having fewer signals because it only deals with partial configurations and does not have to support different configuration modes. For Virtex-II and Virtex-II Pro series, the ICAP furnishes an 8 bit input data bus and an 8 bit output data bus while with the Virtex-4 Series, the ICAP interface has been updated with 32 bit input and output data buses to increase its bandwidth. The ICAP allows the internal logic of the FPGA to reconfigure and to read-back the configuration memory. With combination of either a hard or a soft microprocessor as a controller, dynamic reconfiguration is carried out through the ICAP interface Blodget et al. (2003)

We consider two possible scenarios for dynamically reconfiguring the partial reconfiguration modules: *exo-configuration* and *endo-configuration* as shown in figure 8. The exo-configuration constitutes the traditional way to configuring an FPGA. A configuration bit-stream is controlled by an external processor like DSP. In this way, new modules, or upgraded versions of them, can be created and used at any moment. This approach exhibits upgrade-ability, but the platform is totally dependent on the processor for modifying its function. The endo-reconfiguration, also known as self-reconfiguration, considers a different scenario. An FPGA reconfigures itself using its own local resources. The platform is thus totally independent as it does not require an external source to provide a bit-stream and to decide whether to self-reconfigure. The main draw-back is that partial bit-streams need to be previously generated by a host computer. This approach benefits, thus, of an autonomous reconfiguration with very limited upgrade-ability.



(a) : endo-configuration model

(b) : exo-configuration model

Fig. 6. Exo and Endo FPGA configuration

Our Platform supports both of the previous cited configuration techniques. In fact, at boot time, the initial full configuration bit-stream file is sent to the FPGA. This file includes an internal configuration controller, the internal reconfiguration interface, the initial instantiations of processing bloc units. Depending on the granularity level. One stays internal to the FPGA in case of limited-scale reconfiguration (for co-accelerators configuration) or design parameterization. This implies to interconnect the Micro-blaze to the ICAP internal configuration interface. In this case, small partial bit-streams can be stored inside the FPGA, and the use of endo-configuration lets free the other HW resources of the platform. At a larger scale, configuration for the HW accelerator is external. This implies to interconnect the DSP to the external SelectMap or internal ICAP reconfiguration interfaces. The bit-stream corresponding to the design of HW accelerators are stored in an external SRAM memory.

5. Software development

The great diversity of processing types in a multi-standard application implies a large number of processing configurations to be managed. The configuration management is complex and we believe that a hierarchical approach of configuration control and management could simplify it Delahaye et al. (2005). We combine two configuration features, as presented in section 5.1 and 5.2, in order to create the complete configuration framework for CR testbed.
The proposal of a hierarchical view enables to manage multi-granularity of configurations, which is of particular interest for heterogeneous architectures. The proposed model is composed of three levels of hierarchy detailed in figure 9. A system architecture compliant with this functional model includes one Configuration Manager Unit at level 1 (L1_CMU), several Configuration Manager Units at level 2 (L2_CMU), each of them being responsible for one or several Configuration Manager Units of level 3 (L3_CMU), which directly manage the processing components.

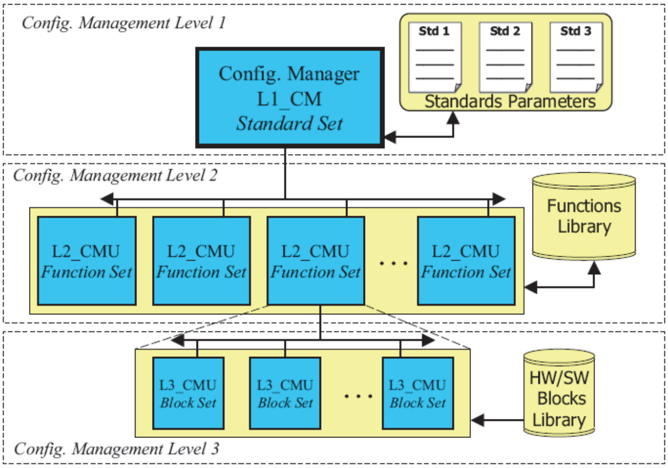


Fig. 7. Hierarchical model of configuration management

5.1 Configuration data-path

As the communication applications are data-flow oriented Delahaye et al. (2005), our approach is based on a data-path model. The functions of the baseband blocks chain are mapped into several Processing Block Units (PBU). Each PBU is optimized using specific reconfigurable hardware resources. In addition, a configuration path, also split into several configuration manager units, controls the reconfigurable processing path. Each CMU, dedicated to a type of PBUs, manages the configuration of a type of baseband function in the chain. The split configuration path offers the possibility to partially reconfigure the baseband chain by an independent reconfiguration of each PBUs.

5.2 Hierarchical management

The hierarchical configuration management model presented in Delahaye et al. (2005) is based on the configuration data path approach. This model is necessary to manage the multi-granularity of configuration required by the different contexts. It is composed of three levels of hierarchy that are detailed below:

- **level 1**: This first high level classification allows a control of category-specific functions to manage parameters at the highest level. The L1_CMU works at the standard level as a host towards the underlying levels of management. This entity is in charge of choosing the functional units which will constitute the entire configuration of the baseband processing chain. At this level, generic functions are handled as generic components. Any hardware implementation is not yet considered.
- **level 2**: The generic functions selected at level 1 are parameterized at the middle level in accordance to standard specifications. The set of attributes of each function is handled by the L2_CMU in order to create each functional context of the entire processing chain.
- **level 3**: The processing data path architecture at this lowest level depends on the reconfigurable computing resources of the hardware architecture. The main task of the L3_CMU in the configuration path is to find the available processing resources and configure them to enable the execution of the functional context created at the middle level.

As presented in figure 8, an hierarchical configuration management is proposed to map processing elements. The CPU and the external storage memories are resources used from a standard PC station. the video coder is implemented in software by a Graphics Processing Unit (GPU). The L1_CMU is a task running on the CPU which controls the configuration of all testbed resources (DSP, PFGA). The DSP works as the master of the (DSP/FPGA) subpart of the platform. The position of master allows the DSP to manage the overall configuration of the functions that run on the SW/HW resources. The FPGA partial reconfigurability is, of course, a mandatory feature to allow reconfiguration of a single component. The L3_CMUs responsible for the configuration of the co-accelerators are implemented as a task into the Micro-blaze soft processor. It allows to perform fine grain reconfiguration of the FPGA without involving any external resource. The hardware and software designs of the processing functions are stored in the external storage memory of the platform where the configuration management can reach them.

6. Application and results

We have designed and implemented a real-time multi-standard application composed of an Active Appearance Models (AAM) schema Cootes et al. (2009) and a digital communication layer (802.11g, UMTS or GSM). The coding application feeds a video coded bit-stream in the transmitter whereas the associated video decoder is connected to the receiver (802.11g, UMTS or GSM). In this section, we give some results of our development approach with the Sundance platform.

6.1 The overall scenario

The platform illustrates the adaptation of the radio link according to the compression of the source in a video-telephony context. A person switches-on his terminal in order to perform an audio-video conversation with another person. At the beginning of the communication,

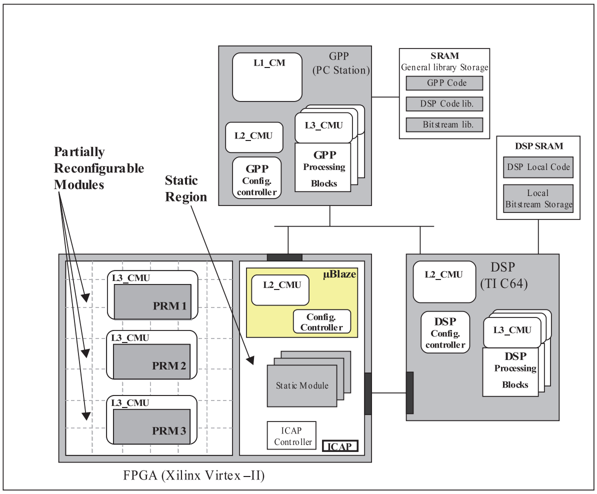


Fig. 8. example of the hierarchical configuration management mapping

the face of the speaker and the background of the image are transmitted using a traditional compression mode. This requires a relatively high data-rate over the time, a model of the person's face is generated at the transmitter's side, and sent to the receiver. Once this model is understood by the receiver, the transmitted parameters of the face's model (orientation, opening of the mouth, of the eyes, direction of the glance) are enough to reproduce the face behavior at the receiver. This permits to save the data amount required to transmit the face of the speaker, by reducing very significantly the data to be transmitted through the air. The data rate variations by step as well as the dynamic reconfigurations of the radio link are illustrated in figure 9.

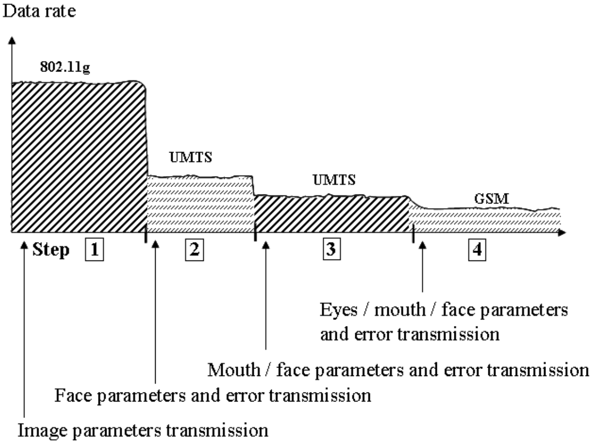


Fig. 9. Standard adaptation as function of the video compression

At the start, the person switches-on his terminal and starts a video-conference service. Video coder starts learning the face model, as well as models for eyes and mouth reconstruction. Then the radio link goes through the following steps.

Step 1: The image is transmitted using a traditional compression mode. The terminal learns the 3D model of the person face and performs a 802.11g modulation with standard error coding.

Step 2: The face model is learned: only high level parameters of the face are transmitted (location, size, orientation) so the receiver can reconstruct the 3D model of the face with its texture on the already sent background. In order to improve the reconstruction at the receiver, errors between the model and the real image are also transmitted by the means of an UMTS modulation with standard error coding.

Step 3: The mouth variations are modeled: The mouth characteristics, as well as high level parameters of the face model are transmitted through UMTS with a very robust error coding on the data for the mouth model.

Step 4: In this last step all face features' models were already learn only high level parameters of all three face, mouth, and eyes models are transmitted, as well as the errors with respect to the real image to help the reconstruction process. GSM modulation with standard error coding can then be used.

The last step is the longer period of the video-call, which permits to reduce very efficiently the global mean throughput necessary for the communication. This justifies the efforts accepted at the beginning of the call in terms of adaptation complexity. Changing from one data rate to another is possible, while permanently reconfiguring the air link characteristics to up a significant degree.

6.2 Experimental results

The matching step of SynDEx consists in performing mapping and scheduling of the algorithm's operations and data transfers onto the architecture processing components and communication media. It is carried out by a heuristic which takes into account durations of computations and inter-component communications to optimize the global application latency.

- Algorithm graph: Application algorithm is represented by a data-flow graph (DFG) to exhibit the potential parallelism between operations. The algorithm model is a direct data dependence graph. An operation is executed as soon as its inputs are available, and this DFG is infinitely repeated. SynDEx includes a hierarchical algorithm representation, conditional statements and iterations of algorithm parts. The application can be described in a hierarchical way by the algorithm graph. The lowest hierarchical level is always composed of indivisible operations. Operations are composed of several input and output ports. Special inputs are used to create conditional statements. Hence an alternative sub-graph is selected for execution according to the conditional entry value. Data dependencies between operations are represented by valued arcs. Each input and output port has to be defined with its length and data type. These lengths are used to express either the total required data amount needed by the operation before starting its computation or the total amount of data generated by the operation on each output port.
- Architecture graph: The architecture is also modeled by a graph, which is a directed graph where the vertices are computation operators (e.g processors, DSP, FPGA) or media (e.g

SHB, SDB, PCI, Ethernet) and the edges are connections between them. So the architecture structure exhibits the actual parallelism between operators. Computation vertices have no internal computation parallelism available. An example is shown in figure 10. In order to perform the graph matching process, computation vertices have to be characterized with algorithm operation execution times. Execution times are determined during the profiling process of the operation. The media are also characterized with the time needed to transmit a given data type.

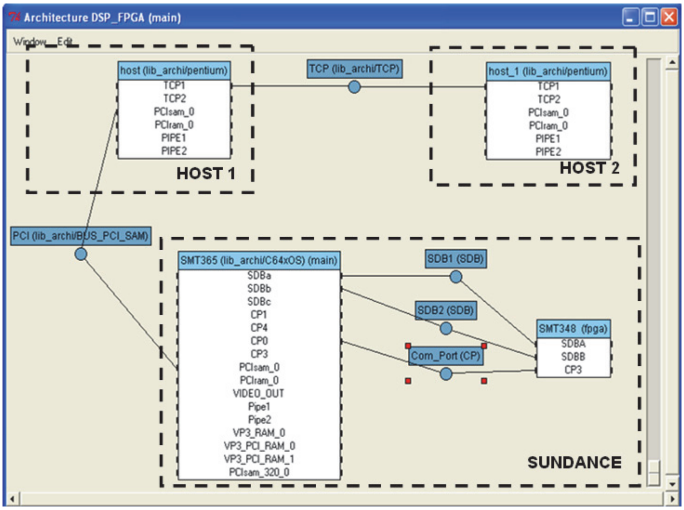


Fig. 10. Syndex model

The output files generated by SynDEx are exploited by our platform to manage correctly the hardware reconfiguration platform. These text files are managed by the L1_CMU of the transmitter (the platform in charge to send video stream) and by L1_CMU of the receiver to be standard compatible with the current or future transmission. In the next section , we present the hardware platform used and its specifications.
This hardware architecture is represented by an architecture graph under SynDEx and in the same manner a DFG is done for the application task (telecommunication chain to be used). Then, the heuristic of SynDEx realizes the adequation between the graph and the hardware and generates constraint files for the DSP and the FPGA that give information for the reconfiguration of the platform. Then the host of the platform sends a new source-code to the DSP and a new order of partial reconfiguration to the FPGA using the architecture described above.

7. Conclusion

In this chapter, we introduce a heterogeneous reconfigurable Sundance platform to support Cognitive Radio in the context of emergency networks. The heterogeneous reconfigurable architecture includes heterogeneous processing elements such as general purpose processors (GPU), DSPs and FPGAs. A key element in this heterogeneous reconfigurable architecture is the run-time partial reconfiguration of the hardware part, which can achieve the reconfigurability in combination with the energy efficiency. A design methodology is needed to map applications onto a heterogeneous platform which has two new features: transaction level modeling of applications and run-time spatial mapping. In the future, we aim to validate the HDCRAM approach, test our PAPR and spectrum sensing algorithms in the case of MIMO system, and sets of algorithms for cognitive radio. The ultimate goal is to build a heterogeneous reconfigurable radio platform to demonstrate the cognitive radio functionalities.

8. References

J. Mitola, "*Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio*," Ph.D. dissertation, Royal Inst. of Tech., Sweden, May 2000.

J. Mitola, G. Maguire, "*Cognitive radio: making software radios more personal, Personal Communications*," IEEE Wireless Communications, Vol. 6, No. 4. (1999), pp. 13-18

A. Nafkha, R. Seguier, J. Palicot, C. Moy, J.P. Delahaye, "A Reconfigurable BaseBand Transmitter for Adaptive Image Coding", *IST Mobile and Wireless Communications Summit*, 1-5 July 2007,

G. Chen, D. Kotz, "*A Survey of Context-Aware Mobile Computing Research*", Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College, Nov. 2000

E. Aarts, H. Harwig, and M. Schuurmans, "*Ambient Intelligence: The Invisible Future*", J. Denning, ed., McGraw Hill, New York, 2001.

X. Qiu, J. Chuang, "Link adaptation in wireless data networks for throughput maximization under retransmissions", in Proceedings of *International Conference on Communications (ICC)*, Vancouver, 1999.

W. Fan, M. Krunz, C. Shuguang; "Price-Based Spectrum Management in Cognitive Radio Networks", *IEEE Journal of Selected Topics in Signal Processing*, vol.2, 2008.

M. Ghozzi, M. Dohler, F. Marx, J. Palicot, "Cognitive radio: methods for the detection of free bands, Towards reconfigurable and cognitive communications", Comptes rendus Physique, Paris, vol. 7, no7, 2006.

F. Jondral, "*Software Defined Radio Enabling Technologies*" edited by W.Tuttlebee, Wiley, 2002.

L.Alaus, J. Palicot, C. Roland, Y. Louet, D.Noguet, "Promising technique of parametrization for reconfigurable radio, the Common Operators Technique : fundamentals and examples", *Signal Processing For Software Defined Radio Handsets*, Springer, 2008.

A-R. Rhiemeier, "Benefits and Limits of Parameterized Channel Coding for Software Radio", 2nd Karlsruhe Workshop on Software Radios, Germany, 2002.

J. Palicot, C. Roland, "FFT: a Basic Function for a Reconfigurable Receiver", *International Conference on Telecommunications*, Papeete, Tahiti, 2003.

V. Rodriguez, C. Moy, J. Palicot, "Install or invoke?: The optimal trade-off between performance and cost in the design of multi-standard reconfigurable radios," Wiley Inter-science, *Wireless Communications and Mobile Computing Journal*, Volume 7 Issue 9, Pages 1143 - 1156, 2007.

J.P. Delahaye, J. Palicot, P. Leray, "A Hierarchical Modeling Approach in Software Defined Radio System Design", IEEE *Workshop on Signal Processing Systems*, Athens (Greece), Nov. 2005.

B. Blodget and S. McMillan and P. Lysaght, "A lightweight approach for embedded reconfiguration of FPGAs," in *Design, Automation and Test in Europe Conference and Exhibition*, 2003.

T. Grandpierre, C. Lavarenne, and Y. Sorel, "Optimized Rapid Prototyping for Real-Time Embedded Heterogeneous Multiprocessors",in CODES, Rome, Italy, May 1999.

J.P. Delahaye, C. Moy, P. Leray, J. Palicot, "Managing Dynamic Partial Reconfiguration on Heterogeneous SDR Platforms", Sdr Forum, November, Los Angeles, USA, 2005.

L. GODARD, C. MOY, J. PALICOT, ”An Executable Meta-Model of a Hierarchical and Distributed Architecture Management for the Design of Cognitive Radio Equipments,” Annals of Telecommunications, Special issue on Cognitive Radio, Volume 64, Numbers 7-8, 2009.

T.F.Cootes, G.J.Edwards and C.J.Taylor, ”Active Appearence Models”, Europeen Conference on Computer Vision,1998.

16

Dealing with VoIP Calls During "Busy Hour" in LTE

Angelos Antonopoulos[1], Elli Kartsakli[2],
Luis Alonso[2] and Christos Verikoukis[1]
[1]Telecommunications Technological Centre of Catalonia (CTTC)
[2]Department of Signal Theory and Communications (TSC),
Technical University of Catalunya (UPC)
Spain

1. Introduction

Long Term Evolution (LTE) is an evolving wireless standard developed by the 3rd Generation Partnership Project (3GPP) which, along with 3GPP HSPA+, 3GPP EDGE Evolution and Mobile WiMAX (IEEE 802.16e), opens the road to 4G technologies. The standard is focused on delivering high data rates for bandwidth-demanding applications and on improving flexibility and spectral efficiency, thus constituting an attractive solution for both end users and mobile operators. An important feature of LTE that differentiates it from conventional mobile standards is the all-IP packet based network architecture, which further ensures the seamless integration of internet applications and facilitates the convergence between fixed and mobile systems.

The radio interface of LTE is based on Orthogonal Frequency Division Multiplexing (OFDM) and supports Multiple-Input-Multiple-Output (MIMO) technology. The standard defines asymmetrical data rates and modulations for uplink and downlink, using different access schemes for each link. In particular, Orthogonal Frequency Division Multiple Access (OFDMA) is employed in the downlink, while the technically similar but less power-demanding Single Carrier – Frequency Division Multiple Access (SC-FDMA) is used in the uplink. In terms of the wireless spectrum allocation, LTE supports variable channel bandwidths that vary from 1.4 to 20 MHz and can be deployed in different frequency bands.

The LTE architecture, referred to as Evolved Packet System (EPS) comprises the Evolved Radio Access Network (E-UTRAN) and the Evolved Packet Core (EPC), illustrated in Fig. 1 (3GPP, 2010). The E-UTRAN consists of a network of enhanced base stations called evolved Nodes B (eNBs) whose main role is to manage the radio resource and mobility in the cell in order to optimize the communication among all User Equipments (UEs). The eNBs can communicate with each other through the X2 interface and can access the EPC by means of the S1 interface. On the other hand, the EPC consists of a control plane node called the Mobility Management Entity (MME) and two user plane nodes, the Serving Gateway and the Packet Data Network Gateway (PDN Gateway or P-GW). These control planes handle the data packet routing within the LTE and towards non-3GPP data networks, respectively.

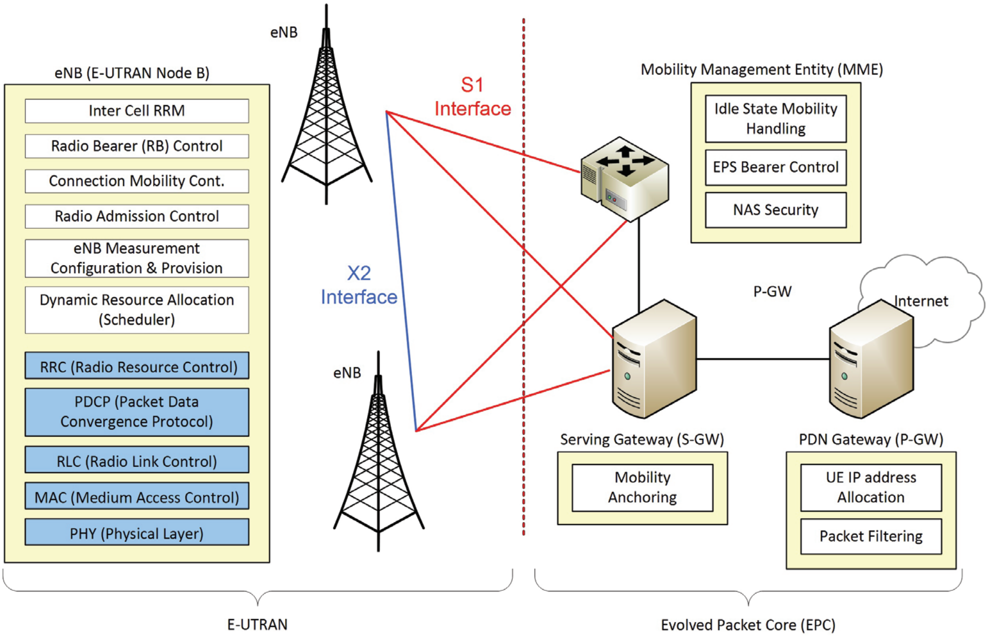

eNB
eNB (E-UTRAN Node B)
Inter Cell RRM
Radio Bearer (RB) Control
Connection Mobility Cont.
Radio Admission Control
eNB Measurement Configuration & Provision
Dynamic Resource Allocation (Scheduler)
RRC (Radio Resource Control)
PDCP (Packet Data Convergence Protocol)
RLC (Radio Link Control)
MAC (Medium Access Control)
PHY (Physical Layer)
S1 Interface
X2 Interface
eNB
Mobility Management Entity (MME)
Idle State Mobility Handling
EPS Bearer Control
NAS Security
P-GW
Internet
Serving Gateway (S-GW)
Mobility Anchoring
PDN Gateway (P-GW)
UE IP address Allocation
Packet Filtering
E-UTRAN
Evolved Packet Core (EPC)

Fig. 1. LTE Architecture and functional split between E-UTRAN and EPC

LTE provides service differentiation by adopting a class-based Quality of Service (QoS) concept. In particular, each data flow between the user equipment and the P-GW (called EPS bearer) is assigned a QoS profile. A total of nine profiles are defined in the specification (3GPP, 2011) that can be mapped to different types of applications such as real time video and voice services, online gaming, etc. Each profile involves the bearer type, the flow priority, an upper bound for the packet delay and the packet error rate. The bearer type indicates whether a Guaranteed Bit Rate (GBR) will be provided to the bearer by permanently allocating network resources during the data session.

The essential difference between GBR and non-GBR bearers is that, in the first case, a connection may be blocked if the network does not have the resources to guarantee the desired QoS of these connections. This concept is known as Call Admission Control (CAC) and it is an important component of radio resource management. CAC algorithms are usually implemented in eNBs and their role is to determine whether a new connection request should be accepted or rejected, depending on the available network resources.

This chapter is focused on CAC policies for handling the admission of Voice over IP (VoIP) calls in an LTE system. Particular interest is laid on the “busy hour” phenomenon, which is defined as the “uninterrupted period of 60 minutes during the day when the traffic offered is the maximum” (Weber, 1968). It is during these intervals of increased traffic that CAC mechanisms play a significant role in the system performance and stability. However, most works in the literature either implement fixed bandwidth reservation schemes or give priority to real time services once they have been admitted to the system. In this chapter, the authors present two different approaches that take into consideration the “busy hour” phenomenon, namely, a dynamic bandwidth reservation scheme and a dynamic CAC mechanism that adapts to the incoming traffic load.

The chapter is organized as follows. Section 2 outlines the related work on admission control schemes for 4G networks found in the literature. The two proposed CAC algorithms are described in Section 3 and their performance is discussed in Section 4. Finally, Section 5 is devoted to conclusions.

2. Related work

Although the admission control concept has been extensively studied in the literature, only a limited number of contributions are developed within the context of 4G networks. The objective of this section is to highlight the recent work on this field in order to provide the reader with an up to date State of the Art.

One of the first attempts towards introducing admission control in the fourth-generation cellular mobile networks has been made by Jeong et al. (Jeong et al., 2005). In their work, the authors present a CAC scheme that supports the QoS requirements of the accepted connections in IEEE 802.16e wireless systems. The objective of the proposed CAC is to maximize the utilization of the resources, considering as basic parameter the capacity estimation of the cell. Furthermore, during the admission and scheduling process, the base station distinguishes the delay-sensitive real-time (RT) from the delay-tolerant non-real-time (NRT) connections. The proposed scheme achieves to fulfil the QoS demands of the connections, but in temporary overloaded situations, only NRT class connections can be admitted, thus excluding entirely the RT traffic.

Qian et al. (Qian et al., 2009) propose a novel radio admission control scheme for multiclass services in LTE systems. The authors introduce an objective function to maximize the number of admitted users and propose a CAC algorithm that implements a service degradation scheme whenever a limitation of resources occurs. In their paper, there is a service differentiation approach, with different portions of bandwidth devoted to each traffic class. However, in the presented numerical results there is no plot that distinguishes the blocking probabilities for the different types of traffic, thus not providing any information about the actual handling of the multiclass services.

Anas et al. propose an admission control algorithm for LTE utilizing the fractional power control (FPC) formula agreed in 3GPP (Anas et al., 2008). In their work, GBR is the only considered QoS of the bearer, while each user is assumed to have a single-bearer. The main idea of their proposed algorithm is that the current resource allocation can be modified in order for the new user to be admitted without violating the power restriction for the physical uplink shared channel (PUSCH).

Lei et al. (Lei et al., 2008) introduce a resource allocation algorithm along with a connection access control scheme for LTE systems with heterogeneous services. Their proposed CAC assigns different portions of bandwidth for real-time and non-real-time connections, thus balancing the ongoing connections of different traffic classes and facilitating the support to potential handoff users. However, the results show that the cell throughput remains the same whether the proposed admission control scheme is applied or not.

In (Kwan et al., 2010) a novel predictive admission control scheme is presented. The authors propose a new cell load measurement method and mechanisms for predicting the load increase due to the acceptance of new connections. In the same content, a resource-estimated CAC algorithm is proposed in (Bae et al., 2009). Specifically, whenever a service request occurs, the resource-estimated CAC algorithm calculates the required amount of resources in order for the request to be served. This amount is determined based on the service type, the modulation and the coding scheme level of the particular user. However, the results show that the proposed CAC is beneficial only in terms of packet delay, since the average data rate and the cell utilization are decreased.

Regarding the bandwidth reservation, a downlink CAC algorithm with look-ahead calls for 3GPP LTE mobile networks is presented in (Sallabi and Shuaib, 2009). The proposed algorithm handles the advance resource reservations, providing a high probability that the advance calls will be immediately served once their session is ready to start. Nevertheless, it is hard to derive useful conclusions since there is no reference or comparison to other admission control methods.

Finally, there are various works on the scheduling phase of LTE that grant priority to the VoIP traffic service (Choi et al., 2007; Puttonen et al., 2008; Saha and Quazi, 2009). The main idea behind these contributions is that the prioritization of the voice packets takes place once the connection has been admitted in the system. This results in higher satisfaction of the VoIP users since the QoS of the voice traffic remains in high levels.

3. Admission control schemes

As mentioned in the introduction, LTE defines a class-based QoS concept, thus providing a simple but effective solution to operators in order to offer differentiation between packet services. Furthermore, recent studies have shown that the proportion of VoIP users show a continuous growth from 28% of users in 2008 (up from 20% of users in 2007) to more than 50% in 2010 (Report Study, 2009). Due to this fact, the proposed schemes focus on voice flows, giving them higher priority comparing to the other types of traffic of the standard.

The problem becomes more intense if we take under consideration the variation of daily traffic volume, where there is a peak during the “busy hour”. In Fig. 2 the mean number of calls per minute to a switching centre taken as an average for periods of 15 minutes during 10 working days (Monday-Friday) is depicted (Iversen, 2010).

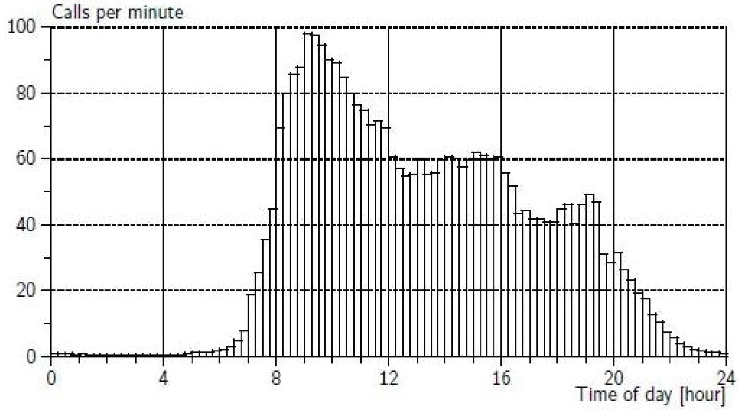


Fig. 2. Typical 24-hour traffic variation

In this section, two CAC algorithms to handle the admission of VoIP calls are presented. The target of the both schemes is to provide enhanced Grade of Service (GoS) to voice traffic flows by improving the acceptance rate of the VoIP calls. Grade of Service is defined as the probability of a call being blocked (BP) or delayed more than a specified interval. From a practical aspect it could be also defined as the probability of a user receiving a network busy signal in a telephone service and can be measured using the following equation:

$$GoS = BP = \frac{Number\_of\_lost\_calls}{Number\_of\_offered\_calls} \tag{1}$$

3.1 Bandwidth reservation-based CAC mechanism (BR CAC)

Our first proposed admission control mechanism is based on the bandwidth reservation concept and is executed under “busy hour” conditions. Under these conditions (i.e. for high arrival rate of VoIP calls), once a connection request arrives at the system, it is mapped onto the corresponding service class. Three main service classes are considered in our scheme: i) the voice GBR ii) the non-voice GBR and iii) the non-GBR traffic types. The two first classes are included in the GBR family, while the third includes the connections that do not require any Guaranteed Bit Rate. In case of voice connections, the request is accepted if the total available bandwidth (BW_T) suffices to serve the incoming connection. On the other hand, restricted bandwidth (BW_T - BW_R) is provided to the other GBR classes, as the algorithm’s aim is to prioritize VoIP calls over other types of connections. In order to deal with the connections that do not require any QoS guarantees (non-GBR), the requests are always admitted, but no bandwidth allocation is considered. The portion of the reserved bandwidth for voice traffic is dynamically changed according to the traffic intensity of the VoIP calls:

$$BW_R = \lfloor \rho_1 \times \beta \rfloor \times BW_1 \tag{2}$$

In the above expression, the traffic intensity ρ_1 is a measure of the average occupancy of the base station during a specified period of time. It is denoted as $\rho = \lambda_1 / \mu_1$, where λ_1 is the mean arrival time for VoIP connections and μ_1 represents their mean service rate (duration). Furthermore, BW_1 is the bandwidth needed for each VoIP call, while $\beta \in [0,1]$ denotes the bandwidth reservation factor.

Formula (2) implies that traffic intensity has an impact on the blocking probabilities of both voice and non-voice connections. It makes sense that applying this bandwidth reservation scheme, the blocking probability for the VoIP connections is decreased, since a portion of bandwidth is exclusively dedicated to this service type. On the contrary, the available bandwidth for the connections of the other service types is decreased and consequently the blocking probability for the specific types increases.

In bandwidth reservation schemes, one of the main difficulties is to avoid the inefficient utilization of system resources. However, in our case, the daily traffic variation establishes the ability to predict an increase in VoIP calls, thus enabling us to tackle this problem. Therefore, our scheme outperforms classic bandwidth reservation mechanisms.

3.1.1 Analytical model

In this section an analytical model for the proposed bandwidth reservation scheme is developed, to derive the blocking probabilities for the different class types. The results are further verified by extensive simulations, presented in the following section.

In order to simplify the analysis, the non-voice connections (e.g. video, data etc) are treated as a single class type with the same characteristics (i.e. arrival rate, bandwidth demand). In this point we must clarify that this simplification takes place only in the admission control process since, after being accepted, the connections are treated according to their different priorities. Furthermore, non-GBR connections are not included in the model as they are always accepted without any QoS guarantees.

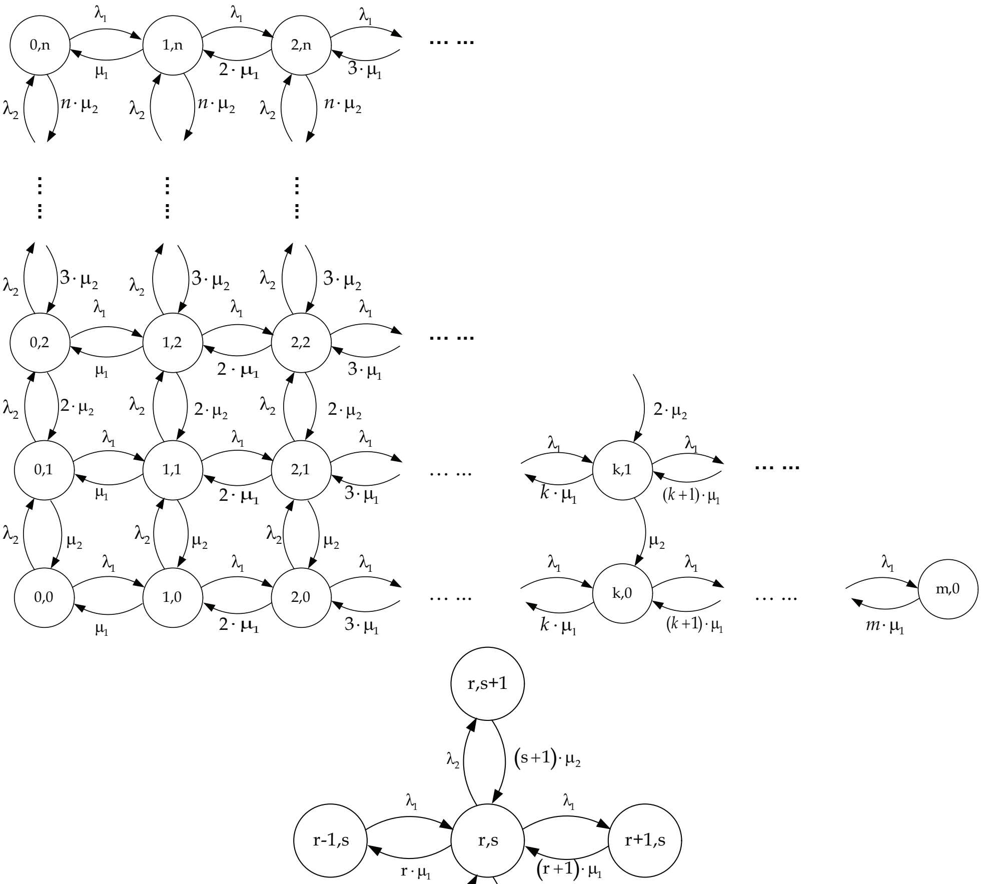


Fig. 3. The two-dimensional Markov model's state transition diagram

Thus, the 2-dimensional continuous Markov model (Fig. 3) can be used to analyze the performance of the proposed scheme. The state space of this Markov model is

$$S=\{(r,s)|0\leq r\leq m,0\leq s\leq n,r\cdot BW_1+s\cdot BW_2\leq BW_T\}\,, \tag{3}$$

where $m = \left\lfloor \frac{BW_T}{BW_1} \right\rfloor$ and $n = \left\lfloor \frac{BW_T - BW_R}{BW_2} \right\rfloor$. The number of VoIP and non-VoIP connections is represented by r and s, respectively. Additionally, BW_T and BW_R represent the overall and the reserved bandwidth, while BW_1 and BW_2 represent the bandwidth that is needed in order to serve each VoIP and non-VoIP connection, respectively. We also define other parameters as follows:

λ_1	Arrival rate of VoIP connections
λ_2	Arrival rate of non-VoIP connections
$1/\mu_1$	Service time for VoIP connections
$1/\mu_2$	Service time for non-VoIP connections

The state transmission diagram of the Markov model is shown in Fig. 3.
Its steady state equation is the following:

$$\begin{aligned} & p_{r,s} \cdot \left(\lambda_1 \cdot \varphi_{r+1,s} + \lambda_2 \cdot \varphi_{r,s+1} \cdot \theta_{r,s+1} + r \cdot \mu_1 \cdot \varphi_{r-1,s} + s \cdot \mu_2 \cdot \varphi_{r,s-1} \right) = \\ & = \lambda_1 \cdot p_{r-1,s} \cdot \varphi_{r-1,s} + \lambda_2 \cdot p_{r,s-1} \cdot \varphi_{r,s-1} \cdot \theta_{r,s} + (r+1) \cdot \mu_1 \cdot p_{r+1,s} \cdot \varphi_{r+1,s} + (s+1) \cdot \mu_2 \cdot p_{r,s+1} \cdot \varphi_{r,s+1} \end{aligned} \tag{4}$$

where $p_{r,s}$ denotes the steady state probability of the system lying in the state (r,s) and $\phi_{r,s}$, $\theta_{r,s}$ denote characteristic functions:

$$\varphi_{r,s} = \begin{cases} 1, & (r,s) \in S \\ 0, & otherwise \end{cases} \tag{5}$$

$$\theta_{r,s} = \begin{cases} 1, & r \cdot BW_1 + s \cdot BW_2 \leq BW_T - BW_R \\ 0, & otherwise \end{cases} \tag{6}$$

The above functions are used in order to prevent a transition into an invalid state, according to the previously defined restrictions. Furthermore, considering the normalization condition $\sum_{(r,s)\in S} p_{r,s} = 1$, the steady state probability for each possible state can be obtained.

The blocking probabilities for VoIP and non-VoIP connections are given by:

$$BP_{VoIP} = \sum_{(r+1)\cdot BW_1 + s\cdot BW_2 > BW_T} p_{r,s} \tag{7}$$

$$BP_{non-VoIP} = \sum_{r\cdot BW_1 + (s+1)\cdot BW_2 > BW_T - BW_R} p_{r,s} \tag{8}$$

3.1.2 Operational example

In order to clarify the mathematical analysis above, we provide two possible states of the system's Markov Chain. Fig. 4 depicts the exact form of the chain in each of the two cases. The first represents the state where there is no available bandwidth for non-voice connections, hence not permitting the transition from *s* to *s+1*. On the other hand, the second represents an equivalent situation along with the assumption that only voice connections are served in the system (*s*=0), thus not allowing the transition from *s* to *s-1* and vice versa.

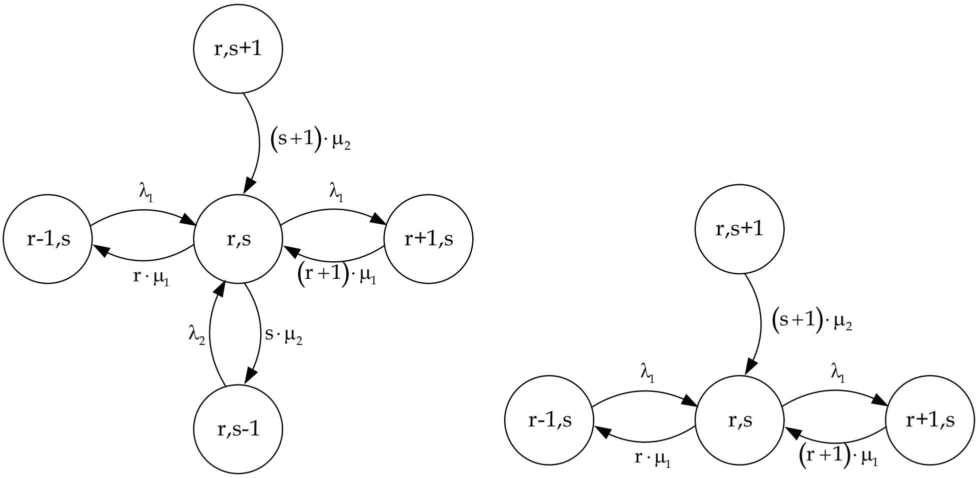


Fig. 4. Two examples of possible states of the system

First case: We assume that the system lies in the state (r, s), subject to the following constraints:

$$\{(r,s),(r+1,s),(r,s+1),(r-1,s),(r,s-1)\} \in S \tag{9}$$

$$r \cdot BW_1 + (s+1) \cdot BW_2 > BW_T - BW_R \tag{10}$$

$$r \cdot BW_1 + s \cdot BW_2 < BW_T - BW_R \tag{11}$$

Under these assumptions and using the definitions of $\phi_{r,s}$ and $\theta_{r,s}$, we derive the steady state equation for the specific case:

$$p_{r,s} \cdot (\lambda_1 + r \cdot \mu_1 + s \cdot \mu_2) = \lambda_1 \cdot p_{r-1,s} + \lambda_2 \cdot p_{r,s-1} + (r+1) \cdot \mu_1 \cdot p_{r+1,s} + (s+1) \cdot \mu_2 \cdot p_{r,s+1} \tag{12}$$

Second case: In this case we assume that the system lies in the state (r,s), subject to the following constraints:

$$\{(r,s),(r+1,s),(r,s+1),(r-1,s)\} \in S \tag{13}$$

$$(r,s-1) \notin S \tag{14}$$

$$r \cdot BW_1 + (s+1) \cdot BW_2 > BW_T - BW_R \tag{15}$$

$$r \cdot BW_1 + s \cdot BW_2 = BW_T - BW_R \tag{16}$$

Considering again the definitions of $\phi_{r,s}$ and $\theta_{r,s}$, we derive the respective steady state equation for this case, that is:

$$p_{r,s} \cdot (\lambda_1 + r \cdot \mu_1) = \lambda_1 \cdot p_{r-1,s} + (r+1) \cdot \mu_1 \cdot p_{r+1,s} + (s+1) \cdot \mu_2 \cdot p_{r,s+1} \tag{17}$$

3.2 Dynamic call admission control algorithm (DCAC)

In the same context, we propose a second CAC algorithm that gives priority to the VoIP calls during the "busy hour". In this scheme, unlike the previous one, no bandwidth reservation takes place, while there is an effort towards a fairer handling of all connections.

According to this CAC scheme, the eNB accepts all the VoIP flows if the available bandwidth suffices in order for the calls to be served. In the case of non-VoIP flows there is an outage probability that depends both on the arrival rate of VoIP requests as well as on the available bandwidth. The requests of non-GBR connections are always admitted, but no bandwidth allocation is considered, since non-GBR flows do not need any QoS guarantees.

The proposed algorithm has two main parameters: the arrival rate of VoIP requests and the available bandwidth of the system. The outage probability for the non-VoIP connections increases either when the arrival rate of the VoIP calls grows or when the available bandwidth decreases. The capacity required in order to serve all the upstream connections can be approximated with the following expression:

$$C_{need} = \sum_{i=1,2} \rho_i \times BW_i \tag{18}$$

All the parameters in the above expression have been already defined. However, it should be stressed that the index i corresponds to different service types and can take values 1 and 2 for VoIP and non-VoIP traffic, respectively.

In case that the system bandwidth suffices to serve the flows of all service types, the outage probability is equal to zero. Due to this fact, the proposed admission control has the same output as classic admission control schemes under light traffic conditions in the network. On the contrary, in overloaded environments where the bandwidth is not sufficient for all connections, an admission control algorithm is required in order to provide different levels of priority to the various connections.

Let us consider the arrival rate of the VoIP requests, defined as λ_1. If this rate is higher than a specific threshold there will be an outage probability for the requests of the other GBR service types. This threshold is defined by the administrator/operator of the network, by considering the network parameters, e.g. the arrival rate of VoIP calls during "busy hour". The value of the outage probability fluctuates between $Pout_{\min}$ and $Pout_{\max}$, depending on the available system bandwidth. In the extreme case that we have no available bandwidth, the overall outage probability becomes $Pout_{\max}$. Adversely, when the total bandwidth of the system is available and no connections are being served, i.e., $BW_{\text{available}}/BW_T = 1$, the outage probability becomes $Pout_{\min}$, since there is enough bandwidth in order for the connections of all types to be served. These borderline values are selected by the system's operator according to each traffic class' desired level of priority. On the other hand, whenever the arrival rate of VoIP connections is smaller than this arrival rate threshold, we assume that we are out of "busy hour" and, therefore, the outage probability equals zero.

The flowchart in Fig. 5 depicts the connection acceptance/rejection procedure in the proposed Dynamic Connection Admission Control (DCAC) algorithm. The basic process of the connection request flow has been described above. In the last part of the algorithm, there is an estimation of the available bandwidth ratio in order to derive the exact value of the outage probability (the higher the ratio, the lower the probability). In particular, the $Pout_{\min}$ is a system parameter, designated by the operator, which determines the desirable level of priority to be assigned to the voice calls. By adding this value to the normalized bandwidth ratio, the outage probability for the specific connection is derived.

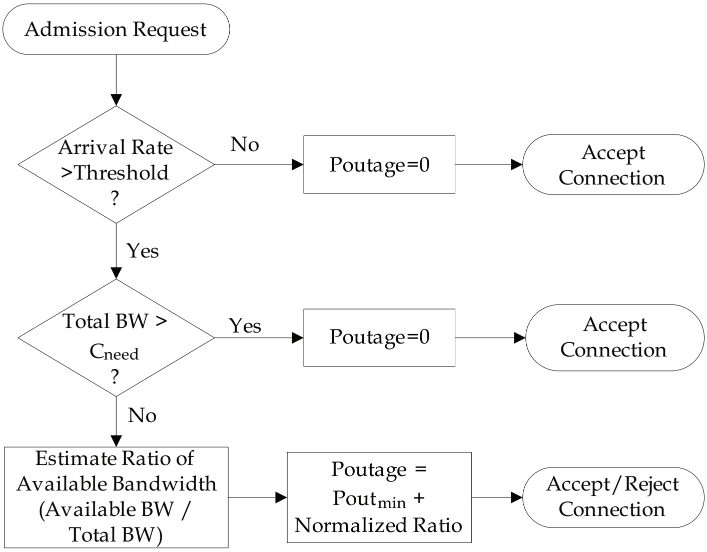

Admission Request
Arrival Rate >Threshold ?
No
Poutage=0
Accept Connection
Yes
Total BW > C_{need} ?
Yes
Poutage=0
Accept Connection
No
Estimate Ratio of Available Bandwidth (Available BW / Total BW)
Poutage = $Pout_{min}$ + Normalized Ratio
Accept/Reject Connection

Fig. 5. Dynamic connection admission control (flowchart)

4. Performance evaluation

In order to evaluate the performance of the proposed CAC schemes and verify the validity of the analytical formulation, corresponding event-driven C++ simulators that execute the rules of the algorithms have been developed. In this section, the simulation set up is described, followed by a discussion of the obtained results.

4.1 Simulation scenario

Based on the physical capabilities of the LTE technology, we assume that the overall bandwidth for the uplink traffic is 4 Mb/s. Assuming that the non-VoIP traffic consists mainly of audio and video data, an average bandwidth of 128 kb/s for each connection is considered (Koenen, 2000). The codec chosen to generate VoIP traffic is the G.711, resulting to a constant bit rate of 64 kb/s. Each result was produced by running the simulation 100 times using different seeds, while we simulate 3600 seconds of real time in order to be in accordance with the definition of “busy hour”.

In order to evaluate the efficiency of the proposed algorithms, a research on the state-of-the-art admission control mechanisms for the LTE standard has been conducted. Several schemes in the literature accept a new connection when the following condition is satisfied:

$$C_{reserved} + TR_i^{service} \leq C_{total} \tag{19}$$

where $C_{reserved}$ represents the capacity reserved by the already admitted connections in the system, $TR_i^{service}$ denotes the traffic rate that should be guaranteed to the new connection i of service type *service* and C_{total} is the total available capacity.

We refer to these methods as capacity-based (CB) algorithms in order to distinguish from our proposed algorithms which are either based on the bandwidth reservation (BR) concept or follow a dynamic approach (DCAC). In order to study the performance of our mechanisms we have carried out simulation tests by varying the VoIP requests arrival rate, thus providing a large range of voice traffic that fluctuates between 15 and 240 connections/min. However, it should be clarified that the rate request of the voice connections remains constant during the busy hour. The system parameters that are presented in Table 1, define that the arrival rate of all connections follows a Poisson distribution, while the mean service time for the connections is exponentially distributed.

Parameter	**Value**
Bandwidth	4 Mb/s
λ_2	Poisson (1 connection/s)
$1/\mu_1$	Exponential (mean 50 s)
$1/\mu_2$	Exponential (mean 50 s)
BW_1	64 kb/s (G.711)
BW_2	128 kb/s
Threshold	0.2 calls/s
β (BR)	1/3
$Pout_{min}$ (DCAC)	0.6
$Pout_{max}$ (DCAC)	0.85

Table 1. System parameters

Under these assumptions and considering λ_1 = 1 connection/s, the system can serve about 98% of the VoIP calls if all the requests of the other classes are rejected, which means that the network is overloaded. Furthermore, in the specific case we use a single admission control based on bandwidth availability (CB) where all the requests are accepted if there is enough bandwidth to serve them, regardless of the class that they belong to, the system serves about 57% of the VoIP flows and 34% of the other flows.

Finally, before proceeding to the simulation results, let us recall that the aim of the proposed schemes is to serve more voice traffic by reducing the GoS, and consequently the blocking probability, of VoIP calls.

4.2 Performance results

Simulation results are compared to those obtained with the mathematical model presented in section 3.1.1. First, it can be observed that the simulation results verify the mathematical analysis, with the difference varying in a range of less than 2% (Fig. 6). Comparing the first proposed admission control to traditional schemes for different values of arrival rates for the VoIP connections, we observe that the BR CAC outperforms single admission control methods in terms of GoS, without any deterioration in the overall system performance. Fig. 6 depicts the GoS among various arrival rates of VoIP calls. It is observed that, using our proposed CAC, a better system performance in terms of voice communication is achieved, as there is a significant enhancement in GoS (10-40%) of VoIP traffic.

On the other hand, the GoS of the other types of connections is increased as expected, but examining the system considering the total number of connection requests (both VoIP and non-VoIP) we achieve a more efficient utilization of system resources as we observe an enhancement in the total GoS ratio for high arrival rates of VoIP connections (i.e. rates greater than 1 connection/s).

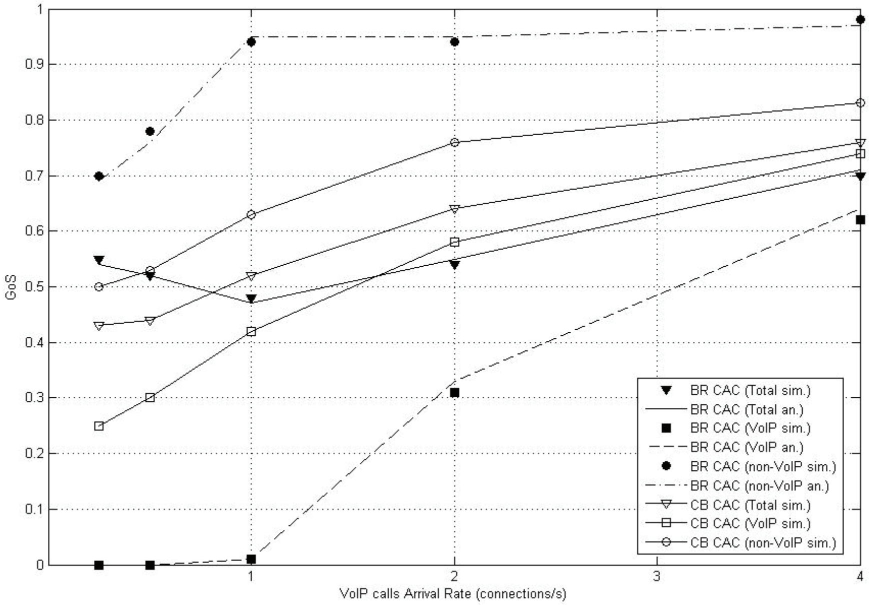


Fig. 6. GoS vs. VoIP Calls Arrival Rate (proposed Bandwidth Reservation (BR) CAC vs. Capacity-based (CB) CAC including analytical results)

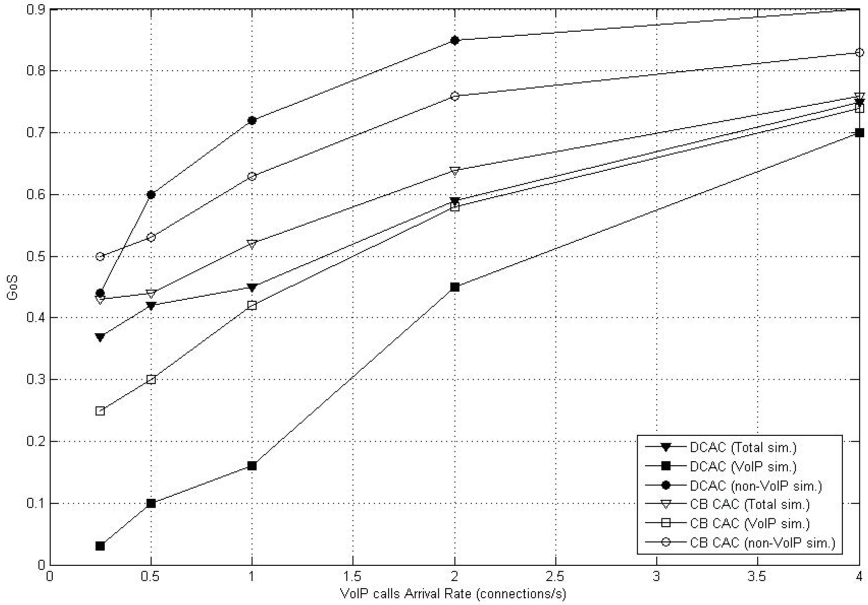


Fig. 7. GoS vs. VoIP Calls Arrival Rate (proposed DCAC vs. Capacity-based (CB) CAC)

The simulation results of the proposed Dynamic Call Admission Control (DCAC) algorithm comparing to the Capacity-based (CB) algorithm are presented in Fig. 7. This algorithm not only improves the voice traffic service, but also enhances the overall system performance. However, in this case the level of prioritization of the VoIP calls over the other type of traffic is lower compared to the bandwidth reservation scenario, thus resulting in a fairer distribution of the system resources.

Furthermore, it is interesting to observe that even for the lower arrival rates of VoIP calls (i.e. 0.25 and 0.5 calls/s) the DCAC handles efficiently the system’s bandwidth, due to its flexibility, while the BR scheme fails to overcome the Capacity-based algorithm. The comparison between the two proposed schemes is given in Fig. 8. In this figure, even if there is no further information provided, it can be clearly seen how the two proposed schemes deal with the different types of traffic, as well as their overall performance. An interesting observation is that, in this particular scenario, the curves for the total GoS for the two schemes cross when the arrival rate is approximately 1.3 connections/s. Below this threshold (i.e. for relatively low traffic conditions) the DCAC outperforms the proposed BR scheme, while above this threshold (i.e. for relatively high traffic conditions) the BR scheme handles the total connections in a more efficient way.

The system’s bandwidth is a main parameter of the DCAC. In Fig. 9 the provided Grade of Service for various values of bandwidth is plotted. As far as networks with restricted bandwidth capabilities are considered, we observe that our proposed dynamic admission control algorithm outperforms single methods, as it improves the GoS of both VoIP calls (11-27%) and of the total number of connections (8-10%) as well.

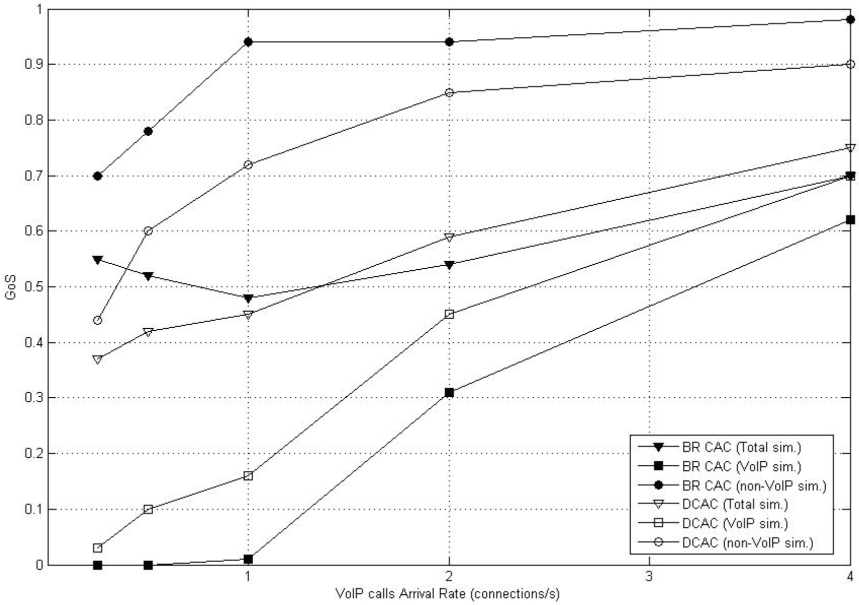


Fig. 8. GoS vs. VoIP Calls Arrival Rate (proposed Bandwidth Reservation (BR) CAC vs. proposed DCAC)

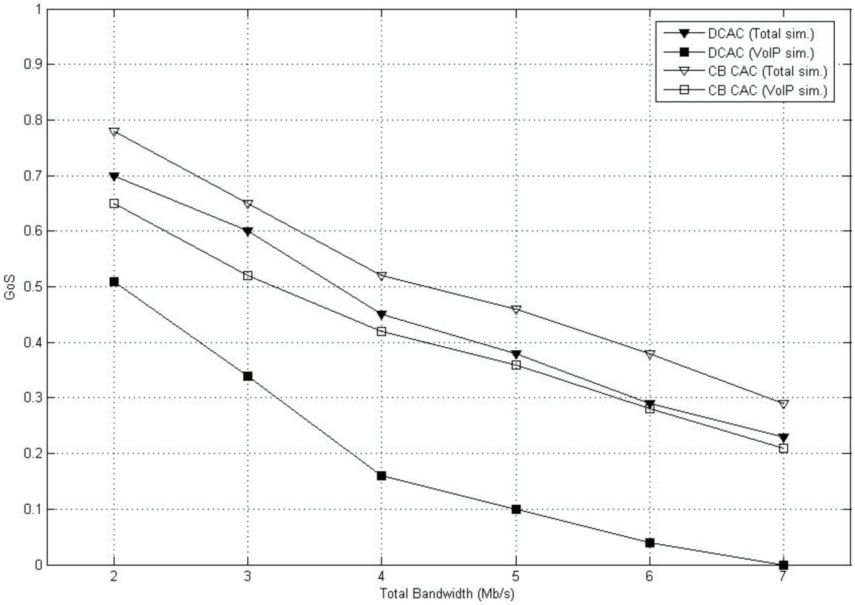


Fig. 9. GoS vs. Total System's Bandwidth (proposed DCAC vs. Capacity-based (CB) CAC)

5. Conclusion

In this chapter, two new admission control schemes for the LTE architecture have been presented. The first mechanism (BR CAC) is based on bandwidth reservation concept, while the second (DCAC) reacts dynamically, depending on the available system's bandwidth. Compared to simple, Capacity-based (CB) admission control methods for 4G networks, the proposed solutions improve the Grade of Service of the voice traffic, without deteriorating the total system performance. The main idea of the proposed schemes is that the base station serves more VoIP calls by considering the "busy hour" phenomenon. Finally, although both the proposed algorithms have been designed with LTE infrastructure in mind, the flexibility of the schemes enables their adaptation to other similar technologies such as IEEE 802.16 (WiMAX).

6. Acknowledgment

This work has been funded by the Research Projects GREENET (PITN-GA-2010-264759), CO2GREEN (TEC2010-20823) and CENTENO (TEC2008-06817-C02-02).

7. References

3GPP (2010). Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) (Release 10); Overall description;

Stage 2, 3rd Generation Partnership Project (3GPP), TS 36.300, v. 10.2.0, Dec. 2010. Available at http://www.3gpp.org/ftp/Specs/html-info/36300.htm

3GPP (2011). Policy and Charging Control Architecture (Release 11) ; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); 3rd Generation Partnership Project (3GPP), TS 23.203, v. 11.0.1, Jan. 2011. Available at http://www.3gpp.org/ftp/Specs/html-info/23203.htm

Anas, M.; Rosa, C.; Calabrese, F.D.; Michaelsen, P.H.; Pedersen, K.I. & Mogensen, P.E. (2008). QoS-Aware Single Cell Admission Control for UTRAN LTE Uplink, *Proceedings of IEEE Vehicular Technology Conference (VTC Spring 2008)*, pp.2487-2491, Marina Bay, Singapore, May 11-14, 2008.

Bae, S. J.; Lee, J. J.; Choi, B. G.; Kwon, S & Chung, M. Y. (2009). A Resource-Estimated Call Admission Control Algorithm in 3GPP LTE System, *Proceedings of ICCSA 2009*, Suwon, Korea.

Choi, S.; Jun, K.; Shin, Y.; Kang, S. & Choi, B. (2007). MAC Scheduling Scheme for VoIP Traffic Service in 3G LTE, *Proceedings of IEEE Vehicular Technology Conference Fall (VTC-2005-Fall)*, pp.1441-1445, Baltimore, USA, Sept. 30-Oct. 3, 2007.

Iversen, V.B. (2010). Teletraffic Engineering Handbook, *Technical University of Denmark*. Available at: http://oldwww.com.dtu.dk/teletraffic/handbook/telenook.pdf

Jeong, S. S.; Han, J. A. & Jeon, W. S. (2005). Adaptive Connection Admission Control Scheme For High Data Rate Mobile Networks, *Proceedings of IEEE Vehicular Technology Conference Fall (VTC-2005-Fall),* vol.4, pp. 2607- 2611, Dallas, Texas, USA, Sept. 25-28, 2005.

Koenen, R. (2000). Coding of Moving Pictures and Audio, ISO/IEC JTC1/SC29/WG11 N4668, March 2000.

Kwan, R.; Arnott, R. & Kubota, M. (2010). On Radio Admission Control for LTE Systems, *Proceedings of Vehicular Technology Conference Fall (VTC 2010-Fall),* pp.1-5, Ottawa, Canada, Sept. 6-9, 2010.

Lei, H.; Yu, M.; Zhao, A.; Chang, Y. & Yang, D (2008). Adaptive Connection Admission Control Algorithm for LTE Systems, *Proceedings of IEEE Vehicular Technology Conference (VTC) 2008,* pp.2336-2340, Marina Bay, Singapore, May 11-14, 2008.

Puttonen, J.; Kolehmainen, N.; Henttonen, T.; Moisio, M. & Rinne, M. (2008). Mixed Traffic Packet Scheduling in UTRAN Long Term Evolution Downlink, *Proceedings of IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2008,* pp.1-5, Cannes, France, Sept. 15-18, 2008.

Qian, M.; Huang, Y.; Shi, J.; Yuan, Y.; Tian, L. & Dutkiewicz, E. (2009). A Novel Radio Admission Control Scheme for Multiclass Services in LTE Systems, *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM) 2009*, pp.1-6, Honolulu, Hawaii, USA, Nov. 30-Dec. 4, 2009.

Report Study (2009). Enterprise VoIP Market Trends 2009-2012, *Osterman Research Inc*., Feb. 2009.

Saha, S. & Quazi, R. (2009). Priority-coupling-a semi-persistent MAC scheduling scheme for VoIP traffic on 3G LTE, *Proceedings of 10th International Conference on Telecommunications (ConTEL 2009), 2009*, pp.325-329, Zagreb, Croatia, June 8-10, 2009.

Sallabi, F. & Khaled Shuaib, K. (2009). Downlink Call Admission Control Algorithm with Look-Ahead Calls for 3GPP LTE Mobile Networks, *Proceedings of IWCMC'09*, Leipzig, Germany, June 21–24, 2009.

Weber, J. (1968). Dictionary of English Language Traffic Terms, , *IEEE Transactions on Communication Technology*, vol.16, no.3, pp.365-369, June 1968.

17

A Semantics-Based Mobile Web Content Transcoding Framework

Chichang Jou
Tamkang University
Taiwan

1. Introduction

With the rapid development of wireless communication technology, in addition to desktop computers, many users are accessing the internet from hand-held appliances, such as tablets, PDAs, and cellular phones. Many new computation paradigms, such as pervasive computing and ubiquitous computing, have been proposed to embrace this emerging portable computation framework. However, because of the difference in users' speed in adopting new technologies, hand-held devices in use have miscellaneous hardware limitations, such as CPU speed, power, memory, bandwidth, and image resolutions. They also have various restrictions in software support, such as operating system, installed programs, real-time processing capability, and rendering functionality. These *ad hoc* limitations have become barriers in human-computer interaction. Most web contents, such as web pages and images, are mainly in the HTML format, which is designed for desktop computers. Without proper modification, most rendered web contents in hand-held devices encounter distorted or fragmented user interface, broken images, slow responses, etc. These ad hoc characteristics of hand-held devices have become a barrier in enhancing web availability.

In this chapter, we illustrate how a semantics-based content adaptation framework could be utilized to fill up the computational gap between mobile devices and desktop computers. The transcoding mechanism of our framework, called Content Adaptation Proxy Server [CAPS], resides behind web servers. In CAPS, web pages and image files are transcoded according to: (1) RDF (Resource Description Framework) (Manola et al., 2004) of web content; and (2) semantics extracted from the CC/PP (Composite Capability Preferences Profiles) (Klyne et al., 2004) client device configuration. These semantic properties will be stored and interpreted inside the Jena Inference System (Carroll et al., 2004) as knowledge facts to obtain proper transcoding parameters for each particular device. Then web pages with proper layout modification and images with proper rendering parameters for each particular device will be constructed and delivered. This technology will recreate contents suitable for resource-limited devices to balance information loss and information availability.

For the rest of this chapter, we will review related work in Section 2. In Section 3, we describe the design principle and system architecture of CAPS. Semantics extraction and knowledge base construction for the Jena Inference System are discussed in Sections 4 and 5.

Section 6 demonstrates web content transcoding process. Section 7 illustrates the implementation of CAPS and demonstrates transcoding examples on various mobile devices. Section 8 concludes this paper.

2. Related work

According how many web content types are handled, web content adaptation could be classified into: universal and specific web content adaptation. Universal web content adaptations are mostly proxy-based and could be applied to web pages and various multimedia types. Their functions include integrating various web content types for the rendering. Section 2.1 will cover proxy-based universal web content adaptation. Specific web content adaptation focuses on the algorithm design. The most frequently studied web content type is HTML files. Due to the complexity in analyzing HTML files, proper adaptation of HTML files in mobile devices is very difficult. Section 2.2 will cover works in using heuristics to adjust layouts of HTML documents to fit a particular mobile device. We will cover how the semantic web technology has been applied to web content adaptation in Section 2.3.

2.1 Proxy-based universal web content transcoding

Nagao et al. (2001) proposed constructing on the Web a system framework using XML and external annotations to Web documents. They proposed three approaches for annotating documents—linguistic, commentary, and multimedia. With annotated documents that computers can understand and process more easily, their framework allowed content to reach a wider audience with minimal overhead.

Lum and Lau (2002) built a quality-of-service oriented decision engine for content adaptation. They designed flows for content negotiation and processing for multimedia contents.

Ardon et al. (2003) prototyped a proxy-based web transcoding system based on network access control, user preferences, and displaying capability of equipments. Since all transcoding procedures were finished in the content provider’s server, this server-centric framework avoided potential copyright problems.

Sacramento et al. (2004) designed the Mobile Collaboration Architecture [MoCA], a middleware for developing and deploying context-aware collaborative applications for mobile users. It comprises client and server APIs, core services for monitoring and inferring the mobile devices' context, and an object-oriented framework for instantiating customized application proxies.

Hua et al. (2006) integrated content adaptation algorithm and content caching strategy for serving dynamic web content in a mobile computing environment. They constructed a testbed to investigate the effectiveness of their design in improving web content readability on small displays, decreasing mobile browsing latency, and reducing wireless bandwidth consumption.

Hsiao et al. (2008) proposed the architecture of versatile transcoding proxy (VTP). The VTP architecture can accept and execute the transcoding preference script provided by the client or the server to transform the corresponding data or protocol according to the user's specification. They adopted the concept of dynamic cache categories and proposed a new replacement algorithm for caches.

Nimmagadda et al. (2010) presented a content adaptation method for multimedia presentations constituting media files with different start times and durations. They performed adaptation based on preferences and temporal constraints specified by authors and generate an order of importance among media files. Their method can automatically generate layouts by computing the locations, start times, and durations of the media files.

2.2 Web page transcoding

Bickmore and Girgensohn (1999) designed a "Digestor System" which was capable of automatic filtering and re-authoring so that WAP-enabled cellular phones could read HTML contents. Their basic idea was to extract plain texts in the HTML document by discarding all formatting elements and unnecessary information. The result was then divided into a navigation page and several plain text sub-pages. They also utilized transcoding cache to diminish the run-time overhead.

Huang and Sundaresan (2000) tried the semantics approach in transcoding web pages to improve web accessibility for users. Their system was designed to improve the interface of e-business transactions and to extend interoperable web forms to mobile devices. They used XML/DTD to specify the semantic and grammatical relationship among web contents, so that web forms could achieve consistency, simplicity and adaptability. The advantage of this system was its ability to provide concept-oriented content adaptation, but it was difficult to be extended.

Buyukkokten et al. (2001) used an "accordion summarization" transcoding strategy where an HTML page could be expanded or shrunk like an accordion. The HTML page was restructured as a tree according to the semantic relationships among its textual sections. All textual sections were split into several Semantic Textual Units, which were automatically summarized. Users could check each summary to expand the node for detailed information. However, this framework only worked in the browser they designed for digital libraries.

Hwang et al. (2003) also treated web page layout as a tree according to the tag hierarchy. They defined a grouping function to transform such a tree into sub-trees, and introduced a filtering mechanism to modify the sub-trees for adequate display in the target device. They analyzed specific web page layout structure and re-authored, according to heuristics, web pages for several mobile devices. Each of their transcoding method could handle only specified layout structures of web pages and did not consider mobile device characteristics.

2.3 Semantic web technologies in web content transcoding

In addition to Huang and Sundaresan (2000), several researchers have tried to incorporate semantic web technologies into web content transcoding. DELI (Butler, 2002), an HP Semantic Lab project, adopted simple negotiation algorithms for rewriting web pages based on context information, like user preference and device capabilities. Due to lack of implementation, its applications were restricted.

Hori et al. (2000) proposed an annotation-based system for Web content transcoding. They introduced a framework of external annotation, in which existing Web documents were associated with content adaptation hints as separate annotation files. This annotation-based transcoding system was then extended with particular focus on the authoring-time integration between a WYSIWYG annotation tool and a transcoding module.

Glover and Davies (2005) used heuristic algorithms to find proper pre-defined web page templates according to device attributes. Their focus was in applying XML/XSLT styles to database contents retrieved in dynamic web pages.

Hsu at el. (2009) proposed a hybrid transcoding approach to combine the traditional transcoding technologies based on ontology-based metadata to improve the rendering problem caused by heterogeneous devices. This heterogeneous markup document transcoding platform was then presented to serve as a transcoding service broker to facilitate interoperability between distributed heterogeneous transcoders.

3. Design principle and system architecture of CAPS

This section illustrates the design principle and system architecture of CAPS.

3.1 Design principle of CAPS

Butler (2001) describes the capabilities of mobile devices in the following categorizations: (1) output: screen, resolution, color, relative size of fonts, sound, etc. (2) input: touchscreen, mouse, keyboard, voice, joystick, etc. (3) processor (4) memory (5) multimedia objects: GIFs, JPGs, WAVs, MP3s, etc. (6) application language: native code, intermediate code. (7) browser language: content markup, client side scripting, applet, and styling. When a particular mobile device receives a web content that it could not handle, it should consider web content adaptation regarding the above aspects. However, from the users' point of view, a more important issue is whether the adapted content could be comprehended. To tackle the above two issues, CAPS follows the following guidelines of the device independence principle of W3C (Gimson et al., 2003):

- For some web content or application to be device independent, it should be possible for a user to obtain a functional user experience associated with its web page identifier via any access mechanism.
- A web page identifier that provides a functional user experience via one access mechanism should also provide a user experience of equivalent functionality via any other access mechanism.
- It should be possible for a user to provide or update any adaptation preferences as part of the delivery context.

3.2 System architecture of CAPS

In this section, we explain the system architecture and the data flows of Content Adaptation Proxy Server [CAPS] in Figure 1. Its adaptation mechanism resides behind web servers. The Proxy Listener is responsible for receiving HTTP requests (message 1) from miscellaneous mobile devices and for dispatching these requests to the Web Content Fetcher. The client's device information as well as user's personal preferences will be embedded inside these requests through CC/PP diff, which is a modified version of predefined CC/PP profile from the hardware manufacturers.

When Proxy Listener accepts a request from a client, it will spawn a working thread in the Web Content Fetcher to handle the request (message 2). Web Content Fetcher performs the standard task of proxy servers. If the requested web content is already in the cache, it will be fetched from Cached Web Content (messages 3.3 and 3.4). If not, then it will be fetched from the source through internet clouds (messages 3.1 and 3.2), and then be saved in the Cached Web Content. The working thread is also responsible for resolving the CC/PP profile diff.

The web content and CC/PP diff will then be sent (message 4) to the Semantics Extractor to acquire the implicit RDF semantic information within the CC/PP diff, HTML web pages and metadata of image files.

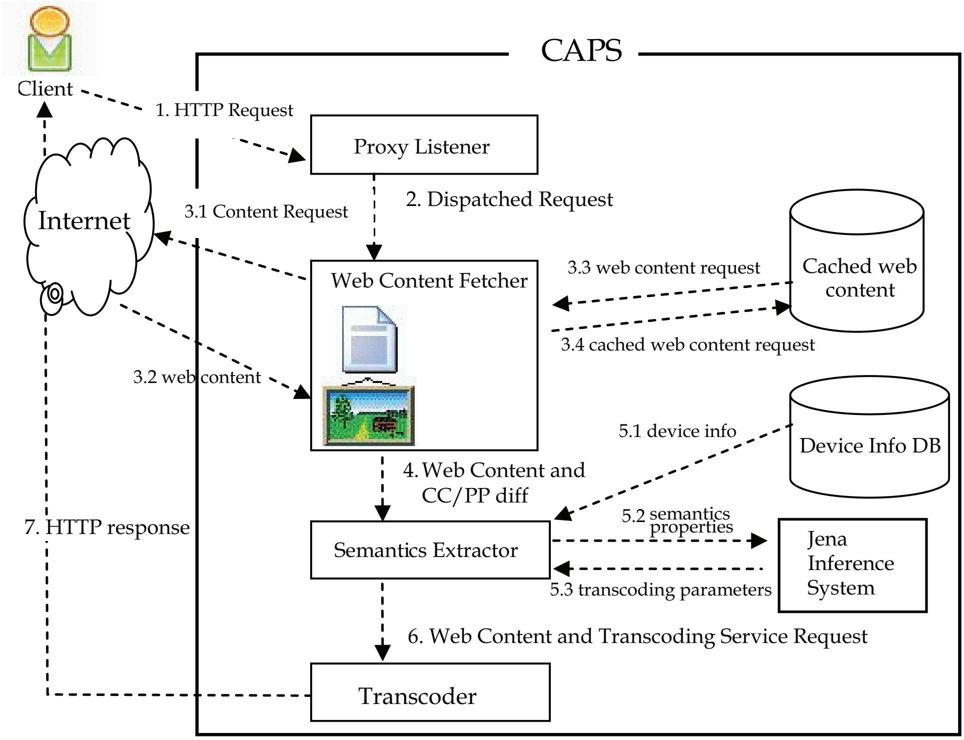


Fig. 1. System architecture of content adaptation proxy server

These semantics information will be sent (messages 5.2) to the Jena Inference System as basic facts. Jena Inference Engine will combine these facts with the knowledge base of CC/PP UAProfile RDFS model, Transcoding Rules and Web Page Auxiliary Vocabulary to determine the proper transcoding parameters in the format of sequential RDF predicates for the requests.

The web content and transcoding parameters will then be passed (message 6) to the Transcoder. Besides the layout rewriting mechanisms, the Transcoder is equipped with transcoding toolkits for image resolution adjustment. The results of the Transcoder processing consist of web pages with modified layout and images with proper resolution and parameters suitable for the requesting mobile device. They will be returned to the client (message 7) and displayed in its browser.

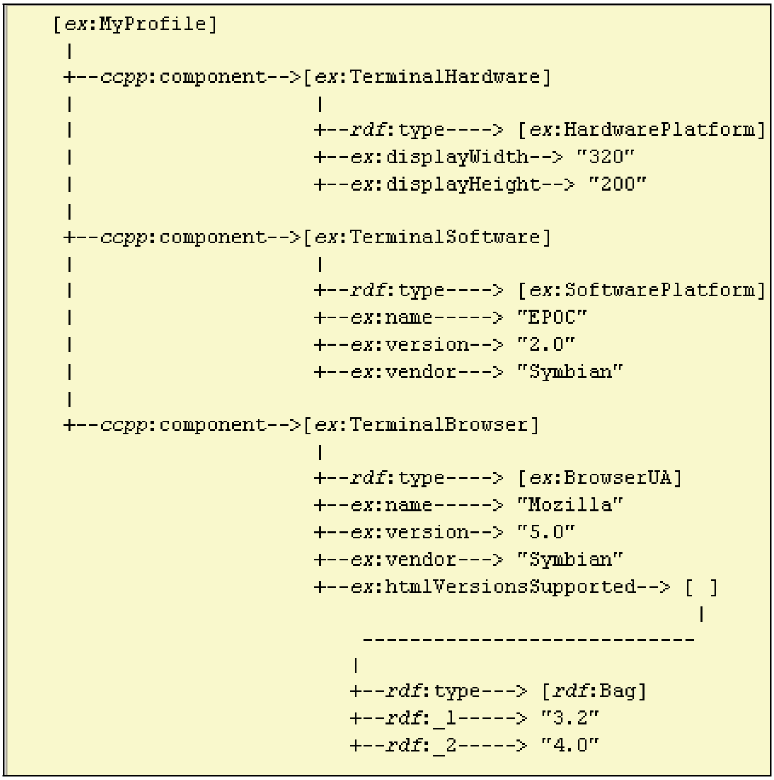
4. Semantics extractor of CAPS

This section illustrates details about semantics extracted from the CC/PP device configurations and the web content. Semantics of device characteristics will be collected through CC/PP diff. The CC/PP semantics of device configurations are RDF compatible already. They could be sent to the Jena Inference System as facts. The total customized device description can be translated into a graph model within the Jena Inference System, which will be described in Section 5.

4.1 Semantics extracted from device configurations

CC/PP is a two-layered user preferences and device capabilities description based on XML/RDF (Ohto & Hjelm, 1999). CC/PP consists of the following three categories: hardware platform, software platform, and browser user agent. Figure 2 shows example detailed attributes in the XML/RDF format (Klyne et al., 2004) for one mobile device.

In CAPS, all predefined CC/PP profiles are stored within a profile directory. A CC/PP diff is manually configured by the user, and is normally used to reflect the user preferences. It is dynamic and can be further modified in later sessions. Many protocols have been proposed to enhance HTTP 1.1 protocol to include CC/PP profile diff. Two of such protocols are CC/PP-ex (Ohto & Hjelm, 1999) and W-HTTP (WAP, 2001). We adopt CC/PP-ex in this framework.



```
[ex:MyProfile]
 |
 +--ccpp:component-->[ex:TerminalHardware]
 |                     |
 |                     +--rdf:type----> [ex:HardwarePlatform]
 |                     +--ex:displayWidth--> "320"
 |                     +--ex:displayHeight--> "200"
 |
 +--ccpp:component-->[ex:TerminalSoftware]
 |                     |
 |                     +--rdf:type----> [ex:SoftwarePlatform]
 |                     +--ex:name-----> "EPOC"
 |                     +--ex:version--> "2.0"
 |                     +--ex:vendor---> "Symbian"
 |
 +--ccpp:component-->[ex:TerminalBrowser]
                       |
                       +--rdf:type----> [ex:BrowserUA]
                       +--ex:name-----> "Mozilla"
                       +--ex:version--> "5.0"
                       +--ex:vendor---> "Symbian"
                       +--ex:htmlVersionsSupported--> [ ]
                                                        |
                          ----------------------------
                          |
                          +--rdf:type---> [rdf:Bag]
                          +--rdf:_1-----> "3.2"
                          +--rdf:_2-----> "4.0"
```

Fig. 2. Example CC/PP for a mobile device

4.2 Semantics extracted from web content

Since most HTML pages are not well-formed, it is hard to extract semantics from them directly. The semantics extractor module first would transform HTML pages into the well-formed XHTML format through the JTidy[1] toolkits. The following two file types of the

[1] JTidy, http://jtidy.sourceforge.net

requested URL will be handled in CAPS: (1) XHTML files: Their metadata are about layouts of the document, possibly with hyperlinks to external textual or binary files. (2) Image files: These are binary files with adjustable parameters, like color depths and resolution. Currently, CAPS could handle JPEG, PNG, and GIF images. For files encoded in indestructible formats, like Java applets, since they could not be adjusted, CAPS would directly forward them to the Transcoder for delivery to the client device.

To extend CAPS to new file types, we need to specify just the metadata about the new file type, and to build the semantics extraction component and transcoding rules for web contents of the new file types.

For XHTML files, the semantic extractor module collects the following schema information: the identification of each XHTML element, and the layout of the XHTML page. We apply XHTML Document Object Model [DOM] (Le Hégaret et al., 2005) tree node scanning to extract node information and relationships among XHTML elements. We solve the element identification problem in an XHTML page by XPath (Clark & DeRose, 1999) so that each node in the DOM tree could be specified accurately.

Statistical or inferred semantics data for the following Web Page Auxiliary Vocabulary will be extracted for each XHTML DOM tree node:

1. **NumberOfWords**: This data indicates number of words in a paragraph. It is used to determine whether the paragraph corresponding to the XHTML node should be split.
2. **NumberOfImages**: This data indicates number of the <IMG> tags in a specific XHTML node. It is used to decide whether a tabular cell is an advertisement banner.
3. **AverageLink**: This is the quotient of the number of links and the number of words within a XHTML node. In web contents with useful information, this value tends to be very high, and all contents in the node should be preserved.
4. **Title**: For XHTML nodes with the <H1> or <H2> tag, or with texts surrounded by pairs of the <B></B> or <STRONG> </STRONG> tags and followed by <BR> immediately, the collected content is treated as a title. This could be used as the title of the sub-page corresponding to this node.
5. **Layout**: This information indicates whether the node is used for layout composition. For example, to determine whether a <TD> is a layout element or an actual tabular cell, we calculate the number of words for the element. If its number of words exceeds a specific threshold, we mark such a <TD> element as a layout element.

Consider the following simplified XHTML page:

```
<HTML>
<BODY>
<TABLE>
    <TR>
      <TD>Gentoo Linux is a totally new linux distribution.
      </TD>
    </TR>
</TABLE>
</BODY>
</HTML>
```

We can describe the <TD> tag in the above page with RDF, XPath and Web Page Auxiliary Vocabulary as follows:

```
<rdf:Description rdf:about="/HTML/BODY/TABLE/TD[1]">
  <rdf:type rdf:resource="html:ELEMENT_NODE" />
  <html:NumberOfWords>8</html:NumberOfWords>
  <html:IsLayout>false</html:IsLayout>
  <html:NumberOfImage>0</html:NumberOfImage>
  <html:NodeName>TD</html:NodeName>
  <html:ChildNodeNumber>0</html:ChildNodeNumber>
  <html:ParentNode rdf:parseType="Resource"
rdf:resource="_:/HTML/BODY/TABLE/TR[1]"/>
</rdf:Description>
```

5. The Jena inference system of CAPS

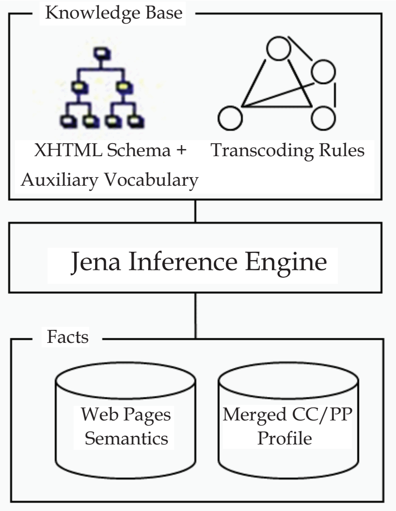


Fig. 3. Architecture of the Jena inference system

We use Jena (Carroll et al., 2004), a semantic web toolkit of “Device Independent” ideal, to determine the transcoding parameters, which are represented as a sequence of RDF predicates. The Jena Inference System, displayed in Figure 3, has three main components. These components are utilized in CAPS as follows:

1. **Knowledge Base** – It contains the acquired knowledge and rules in deciding the content adaptation parameters. XHTML schema is derived by mapping from XHTML XML schema to RDF/RDFS as one knowledge base. Transcoding Rules contain rules using web content ontology and device characteristics ontology for transcoding. Auxiliary Vocabulary for transcoding parameter decision are also described by RDF/RDFS and serialized into Jena knowledge base. All RDF knowledge is serialized in the XML format to provide more flexibility and interoperability in content adaptation.
2. **Jena Inference Engine** – This is the decision engine to inference and to generate transcoding parameters. We make use of the engine without any modification.

3. **Facts** – These are facts supported in the form of instantiated predicates. In CAPS, they are the semantic data collected by the Semantic Extractor.

The transcoding rules of CAPS are to link the web content description and device information in CC/PP, which are transmitted via CCPP diff. In other words, these rules would map the user's or device's requirements into parameter settings of web content. We follow the semantic network model (Hayes & McBride, 2004) defined by W3C, and define the following sets of facts to explain formal meaning of the transcoding rules:

1. **D**: CC/PP Facts transmitted from the client side. These are used to describe characteristics of devices or user preferences.
2. **C$_t$**: All web content semantics for MIME type t obtained by "semantics extractor".
3. **L**: All constraints and limitations of web content. For example, image width less than screen width, image color depth less than or equal to 24 bit, etc.

In addition, we define several vocabularies to represent the actions in removing an element in text/html document, and in changing the coding of a web page. For MIME type t, suppose the set of transcoding operations for type t is **O$_t$**. The union of the above sets forms the set of all statements of our RDF model:

$$S_{RDF} = \bigcup_t (C_t \cup O_t) \cup L \cup D$$

The preconditions of the transcoding rules is a subset of $\bigcup_t C_t \cup L \cup D$. The result of these transcoding rules is a sequence of operations. An example sequence is: update image size, change coding, etc. Suppose the set of possible transcoding actions for MIME type t is called **M$_t$**. In CAPS, **M$_t$** is formally defined as:

$$M_t = O_t \cup C_t$$

We call **M$_t$** the transcoding module for MIME type t. The result of transcoding rules for web content with type t could be represented a subset of **M$_t$**. Finally, the set of all possible transcoding rules for MIME type t could be defined as:

$$R_t = P(C_t \cup D \cup L) \times P(M_t)$$

where P( ) is the power set function. Thus, the input of the transcoding rules is a set of statements about web content (**C$_t$**), device CC/PP configurations (**D**), and constraints (**L**). The output of the transcoding rules then is a member of P(**M$_t$**) that are the required actions for the transcoding.

We apply heuristics to design the transcoding rules in the "IF…THEN…" format. If the precedent parts of a rule are all true, then the consequent part of the rule would be added as a statement of transcoding parameters into the knowledge base. We provide rules regarding device characteristics by defining restriction rules using first order predicate logic. Rules are categorized to back up each other. For example, if rules for HP 6530 PDA are not sufficient, then rules for PPC Pocket PC could be used. The following example rule is to reduce the width of a JPG image file to fit the screen width of the mobile device:

Accept(DV, "image/jpeg") ^ Format(I, "image/jpeg") ^ ScreenWidth(DV,DW) ^ ImageWidth(I, IW) ^ LessThan(DW, IW) → ScaleImageByWidth(I, DW)

In the above RDF rule, DV, I, DW, IW are variables for device, image, device width, and image width. Accept(DV, “image/jpeg”) is a statement of the RDF model to express that the device could render the multimedia type “image/jpeg”. The other predicates before the “→“ symbol could be easily interpreted similarly. So the predicates before the “→“ symbol are to check the conditions about web content (Format and ImageWidth), device CC/PP configurations(Accept), and constraints(LessThan). The predicate after the “→“symbol indicate the action to be performed: ScaleImageByWidth. The above inference rule could be expressed by the following Jena rule:

```
[ ScaleImageByWidth:
(system:Content content:Width ?image_width),
(system:HardwarePlatform ccpp:DisplayWidth ?display_width),
    lessThan( ?display_width, ?image_width ) ->
(system:Content content:ScaleImageByWidth ?display_width) ]
```

The above rule is named ScaleImageByWidth. In Jena rules, names prefixed with ‘?’ are variables. The namespaces “system”, “ccpp”, and “Content” point to the content adaptation proxy server, the standard UAProf Schema (WAP, 2001), and web content under transcoding, respectively. The above rule means that if the width of the device (?dispaly_width) is less than the width of the image (?image_width), then set width of the image as width of the device, and set height of the image proportionally with respect to the adjustment ratio of the width.

6. Transcoder of CAPS

The Transcoder is composed of several transcoding modules corresponding to file types of XHTML, JPEG, etc. It could be extended to handle other file types. According to the transcoding actions and parameters produced by Jena, it performs detailed content adjustment and filtering. For image files, currently the system not only transforms image files into the same format with different parameters, but also transforms image files into different formats.

According to file type of the web content, the Transcoder dispatches them to the corresponding adaptation component. To perform the required content transcoding operations, it will query the inferred RDF model through Jena’s RDF query language RDQL to obtain the required transcoding parameters. The use of RDQL could prevent tight coupling of the transcoding components. An example RDQL query is demonstrated as follows:

```
SELECT ?predicate, ?object
WHERE ( system:Content, ?predicate, ?object)
USING system FOR http://www.im.abc.edu/~def/proxy.rdfs#
```

USING is to specify the name space. This query could obtain all RDF statements with subject system:Content, where system is the name space and Content represents web content currently under transcoding. Transcoding query results are represented as instantiated predicates. For example, if the transcoding predicate for an image file is ScaleImageByWidth, then the Transcoder would adjust the image width and height proportionally. After all RDQL query results are handled, the resulting web content would be returned to the client.

7. Implementation of CAPS

We implement the content adaptation proxy server in the Fedora Linux 13 operating system by the Java Language J2SDK 1.6.0. We use the package org.w3c.dom for handling XHTML DOM trees of the web pages, and use the java.awt.image.BufferedImage package for handling the JPEG, PNG, and GIF image files.

7.1 Implementation details

The Semantics Extractor is implemented by a Java interface, a factory for semantic extraction, and one class for each supported file type. The obtained semantic attributes for the implemented MIME types are listed in Table 1.

MIME type	Semantic attributes extracted
image/jpeg	Image Height Image Width Image Color Depth Image Format
image/png	
image/gif	
text/html	Encoding of the web page Number Of Words In line Document Type

Table 1. The extracted Web content attributes in CAPS

The transcoding rules for images are listed in Table 2, and some of the transcoding rules for HTML pages are listed in Table 3.

The transcoding modules are responsible for receiving the transcoding parameters and performing the actual content adaptation. In CAPS, we have implemented modules for HTML, JPEG, PNG, and GIF files. The JPEG, PNG, and GIF image files are handled by the java.awt.image.BufferedImage package, while the HTML files are handled by the org.w3c.dom package. The web content extraction and parsing component obtains the requested content from the internet, and use JTidy to reformat the web page into the XHTML format and then build the DOM tree for the page. Most of the transcoding modules are implemented by built-in Java classes. The only module that we do use non-built-in Java classes are for the transcoding of images to ASCII files, which is completed through open source tool Asciizer[2].

7.2 System test of CAPS

To demonstrate the functionalities of this framework, we tested three client mobile devices: HP iPAQ hx2400, Symbian S80 Simulator, and Panasonic EB-X700. We would like to show the effect of the following two CC/PP parameters: supported file types and display size. The goal of the adaptation is to avoid the use of horizontal scroll bar, so as to increase the readability of transcoded pages and images. The related specifications and restrictions of these devices are listed in Table 4.

[2] Asciizer, http://asciizer.sourceforge.net

Transcoding Rule	Comment
[ScalingImageByWidth: (system:Content content:Width ?image_width), *(system:HardwarePlatform ccpp:DisplayWidth ?display_width),* *lessThan(?display_width, ?image_width) ->* *(system:Content content:ScaleImageByWidth ?display_width),* *(system:ScaleImageByWidth system:TranscodeType "text/jpeg")]*	Adjust image width to fit in screen size
[ExtractColor: *(system:Content content:Width ?image_color_depth),* *(system:HardwarePlatform ccpp:ScreenColorDepth ?device_color_depth),* *lessThan(?device_color_depth, ?image_color_depth)* *->(system:Content content:ReduceColorDepth ?device_color_depth),* *(system:ReduceColorDepth system:TranscodeType "text/jpeg")]*	Modify image color depth to match the display capability of the device.
[PngToJpeg: *(system:Content content:Type "image/png"),* *(system:SoftwarePlatform ccpp:CcppAccept ?Bag),* *noValue(?Bag ?li "image/png"),* *(?Bag ?li "image/jpeg")* *-> (system:Content system:TransformTo "image/jpeg"),* *(system:TransformTo system:TranscodeType "text/jpeg")]*	Transform PNG files to JPEG.
[JpegToPlainText: *(system:Content content:Type "image/jpeg"),* *(system:SoftwarePlatform ccpp:CcppAccept ?Bag),* *noValue(?Bag ?li "image/jpeg"),* *(?Bag ?li "text/plain")* *-> (system:Content system:TransformTo "text/plain"),* *(system:TransformTo system:TranscodeType "text/jpeg")]*	Transform JPEG to plain text.
[JpegToHTML: (system:Content content:Type "image/jpeg"), *(system:SoftwarePlatform ccpp:CcppAccept ?Bag),* *noValue(?Bag ?li "image/jpeg"), (?Bag ?li "text/html")* *-> (system:Content system:TransformTo "text/html"),* *(system:TransformTo system:TranscodeType "text/jpeg")]*	Transform JPEG to HTML.

Table 2. Transcoding rules for image files

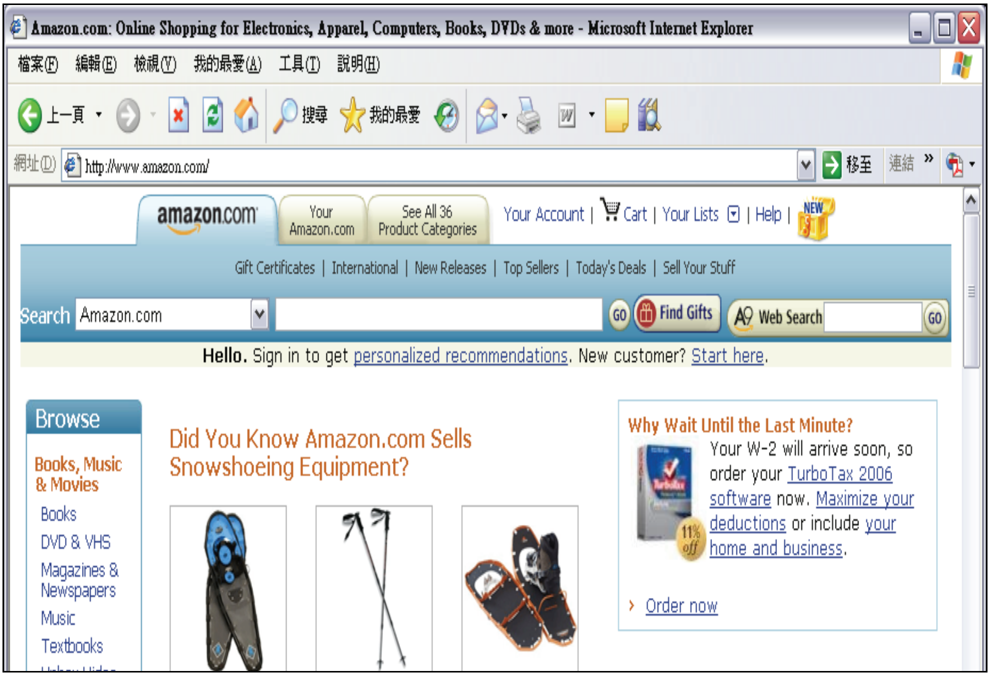
Figure 4 shows the upper part of the tested web page (http://www.amazon.com) in a Microsoft IE 8 browser in a desktop computer. The upper parts of the transcoded pages in the built-in browser for the three tested mobile devices are displayed by two screen shots in Figures 5 to 7. All resulting transcoded web pages satisfy the goal of avoiding the use of horizontal scroll bar by adjusting the page layout, image size, and image resolution. Unsupported CSS, Javascripts, flashes, div's and tables are filtered out.

Transcoding Rule	Comment
[ExtractTableContent: *(?node content:NodeName "table"),* *(system:BrowserUA ccpp:TablesCapable "No")* *->(system:Content system:ExtractTableContent ?node) ,* *(system:ExtractTableContent system:TranscodeType "text/html")]*	If the browser in the mobile device does not support table, then extract the content.
[FilterCSSScript: *(?node content:NodeName "style"),* *(system:BrowserUA ccpp:StyleSheetCapable "No")* *-> (system:Content system:RemoveNode ?node)]* *[FilterCSSScript:* *(?node content:NodeName "style"),* *(system:SoftwarePlatform ccpp:CcppAccept ?Bag),* *noValue(?Bag ?li "text/css")* *-> (system:Content system:RemoveNode ?node) ,* *(system:RemoveNode system:TranscodeType "text/html")]*	If the browser in the mobile device does not support CSS, then filter the CSS content.
[FilterFlash: *(?node content:NodeName "object"),* *(system:SoftwarePlatform ccpp:CcppAccept ?Bag),* *noValue(?Bag ?li "x-application/flash"),* *(?node content:InlineDocumentType "x-application/flash")* *-> (system:Content system:RemoveNode ?node) ,* *(system:RemoveNode system:TranscodeType "text/html")]*	If the browser in the mobile device does not support Flash, then filter the Flash content.
[TransformToPlainText: *(system:SoftwarePlatform ccpp:CcppAccept ?Bag),* *noValue(?Bag ?li "text/html"), (?Bag ?li "text/plain")* *-> (system:Content system:TransformTo "text/plain"),* *(system:TransformTo system:TranscodeType "text/html")]*	Transform HTML to plain text.

Table 3. Transcoding rules for HTML files

Device	HP iPAQ hx2400	Symbian S80	Panasonic EB-X700
Category	PDA	Smart Phone	Smart Phone
Operating System	Windows Mobile 5.0	Symbian Series80	Symbian Series60
Browser	IE Mobile	Built in	Built in
Supported file types	text/html text/css image/jpeg image/png image/gif	text/xhtml text/css image/jpeg image/png image/gif	text/chtml image/jpeg
Display size	480 x 320 (pixels)	220 x 640 (pixels)	176 x 148 (pixels)
Connection	Bluetooth	WLAN	GPRS

Table 4. Specifications and restrictions of tested mobile devices


Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & more - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
網址(D)
http://www.amazon.com/
移至
連結
amazon.com
Your Amazon.com
See All 36 Product Categories
Your Account | Cart | Your Lists | Help |
NEW
Gift Certificates | International | New Releases | Top Sellers | Today's Deals | Sell Your Stuff
Search
Amazon.com
GO
Find Gifts
Web Search
GO
Browse
Books, Music & Movies
Books
DVD & VHS
Magazines & Newspapers
Music
Textbooks
11% off
Order now

Hello. Sign in to get personalized recommendations. New customer? Start here.

Did You Know Amazon.com Sells Snowshoeing Equipment?

Why Wait Until the Last Minute?

Your W-2 will arrive soon, so order your TurboTax 2006 software now. Maximize your deductions or include your home and business.

Fig. 4. Test web page in a desktop computer

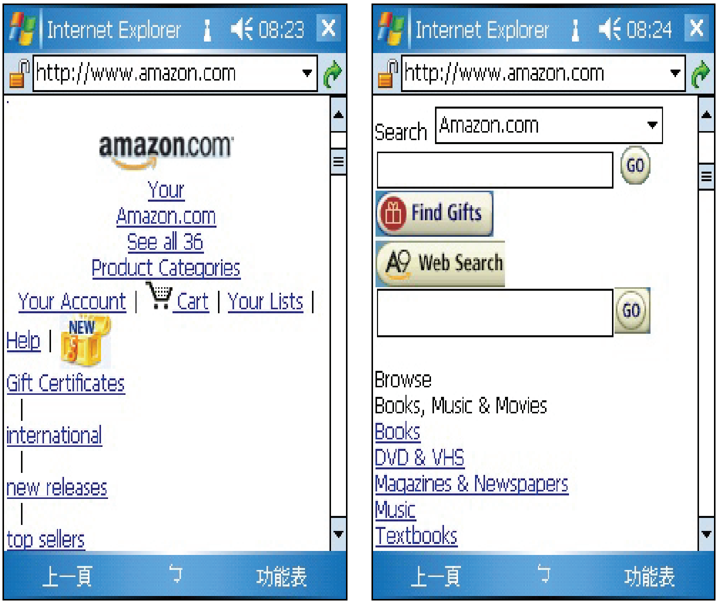


Fig. 5. Results of the test page in HP iPAQ hx2400

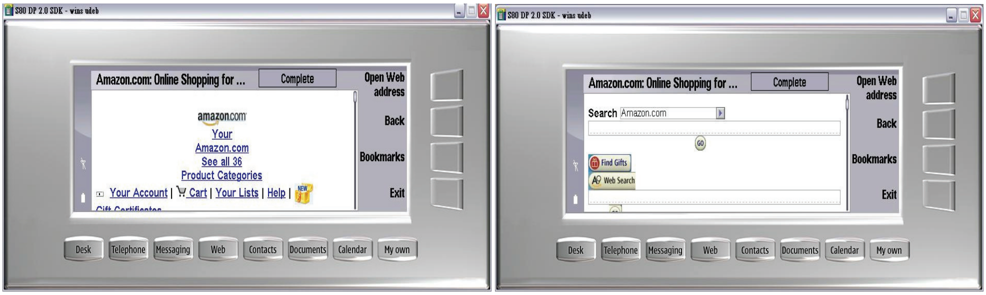


Fig. 6. Results of the test page in Symbian Series80

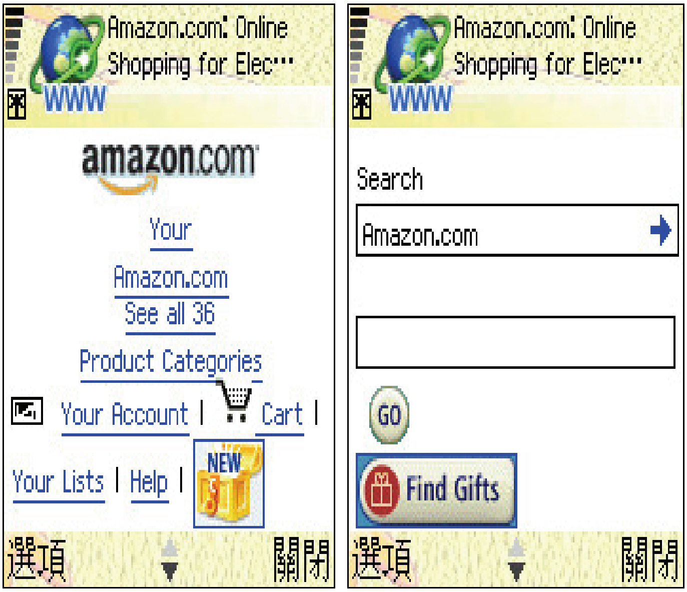


Fig. 7. Results of the test page in Panasonic EB-X700

7.3 Comparison of CAPS with related work

In Table 5, we compare CAPS with related works mentioned in Section 2.1 in the following aspects: purpose, implementation levels, server deployment, dynamic adaptor loading, and transcoding parameter selection method. Some results are from the comparison about proxy-based web content adaptation system in Endler et al. (2005).

In the aspect of purpose, systems focused on Quality-of-Service [QoS] would emphasize on the reduction of transmission time and on improving users' browsing experiences. Systems focused on multimedia would emphasize on the handling of multimedia files, like summary of sound and images, and reduction of sampling frequencies. Systems with general purpose would have a flexible framework and emphasize on the dynamic deployment capability.

To handle the fast growing number of accepted web content types in mobile devices, web content adaptation systems normally would be equipped with the dynamic deployment of transcoding modules to improve the scalability of the system. Only MARCH and CAPS provide dynamic adaptor loading of transcoding modules. In MARCH, the transcoding server was dynamically determined by the dynamic transcoding path, where each node represents a transcoding server. After traversing all nodes in the path through message passing, in which web content is encapsulated in HTTP/1.1, the web content transcoding is also finished. In CAPS, the dynamic adapter loading is implemented through Java's dynamic binding and Run-Time Type Identification [RTTI]. The transcoding modules are decided on-the-fly and thus require less loading than MARCH.

Systems	MARCH (Ardon et al.)	Lum and Lau (2002)	MoCA (Sacramento et al., 2004)	Nagao et al. (2001)	CAPS (this study)
Purpose	universal	QoS for multimedia	universal, caching	multimedia	universal
Implementa tion level	application	-	middleware	application	application
Server deployment	server-driven, proxy-based	-	server-driven	proxy-based	proxy-based
Dynamic adapter loading	yes	no	yes	no	yes
Transcoding parameter selection method	rules	heuristics	rules	heuristics	rules

Table 5. Comparison of CAPS with related work

In parameter selection, there are two approaches: heuristics and rules. The benefits of heuristics-controlled parameter selection are easy to be implemented in programs and faster response time. It drawbacks include less number of handled cases, and less flexibility. Once logic used in parameter selection is changed, the programs must be re-compiled. On the other hand, use of rules in parameter selection provides better flexibility. Additionally, with the capability of reasoning, more transcoding strategies could be obtained than heuristics. However, the drawback of rules-controlled parameter selection is the difficulty to write correct and proper rules.

CAPS uses light-weight transcoding components through flexible API. It could be easily deployed as distributed processing by RMI or Web Services. Therefore, the costs for deployment, allocation, and scalability could be greatly reduced.

8. Conclusion

We designed and implemented a flexible and robust framewrok, called CAPS, for web content adaptation using RDF semantics from CC/PP device characteristics, XHTML web pages, and JPEG, PNG, and GIF image files. Past researches in this area either did not take device characteristics into consideration, or were not a general purpose solution for miscellaneous mobile devices. We made use of the Jena Inference System to obtain the transcoding parameters through the fact and knowledge base built from the collected semantics.

In CAPS, a single copy of the web pages could serve many different mobile devices. Previous tedious web page rewriting labour for mobile devices could be saved. Our framework could be easily extended to new file types by importing related semantics and transcoding modules.

We plan to incorporate support of style sheet and web form specifications, such as CSS/Mobile (Schubert & Berjon, 2008) and XHTML for Mobile (McCarron et al., 2010), into this semantics-based content adaptation framework. By supporting these dynamic web pages, the increased user interactivity would accelerate user acceptance of pervasive computing.

9. Acknowledgment

This work was supported in part by Taiwan's National Science Council under Grant NSC 95-2221-E-032 -027.

10. References

Ardon S.; Gunningberg P.; Landfeldt B.; Ismailov Y.; Portmann M. & Seneviratne A. (2003). MARCH: a distributed content adaptation architecture. *International Journal of Communication Systems*, Vol. 16, No. 1, pp. 97-115, ISSN 1074-5351

Bickmore T. & Girgensohn A. (1999). Web page filtering and re-authoring for mobile users. *The Computer Journal*, Vol. 42, No. 6, pp. 534-546, ISSN 1460-2067

Butler M. (2001). Current Technologies for Device Independence, *HP Laboratories Technical Report*, No. 83

Butler M. (2002). DELI: A Delivery context library for CC/PP and UAProf, *External technical report,* HP Semantic Lab

Buyukkokten O.; Garcia-Molina H. & Paepcke A. (2001). Accordion summarization for end-game browsing on PDAs and cellular phones, *Proceeding Of the 2001 SIGCHI Conference on Human Factors in Computing Systems*, pp. 213-220, ISBN 1-58113-327-8

Carroll J.; Dickinson I.; Dollin C.; Reynolds D.; Seaborne A. & Wilkinson K. (2004). Jena: implementing the semantic web recommendations, *Proceedings of the 13th World Wide Web Conference*, pp. 74-83, ISBN 1-58113-844-X

Clark J. & DeRose S. (1999). XML Path Language (XPath) v. 1.0, W3C, Available from http://www.w3.org/TR/xpath

Endler M.; Rubinsztejn H.; Rocha R. C. A.; Sacramento V. (2005). Proxy-based Adaptation for Mobile Computing, *Technical Report*, Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Available from http://ftp.inf.puc-rio.br/pub/docs/techreports/05_24_endler.pdf

Gimson R.; Finkelstein S. R.; Maes S. & Suryanarayana L. (2003). Device Independence Principles, W3C Working Group Note, Available from http://www.w3.org/TR/di-princ/

Glover T. & Davies J. (2005). Integrating device independence and user profiles on the web. *BT Technology Journal*, Vol. 23, No.3, pp. 239-248, ISSN 1358-3948

Hayes P. & McBride B. (2004). RDF Semantics, W3C Recommendation, Available from http://www.w3.org/TR/rdf-mt/

Hori M.; Kondoh G.; Ono K.; Hirose S. & Singhal S. (2000). Annotation-based web content transcoding, *Computer Networks*, Vol. 33, No. 1-6, pp. 197-211, ISSN 1389-1286

Hsiao J. L.; Hung H. P. & Chen M. S. (2008). Versatile transcoding proxy for internet content adaptation. *IEEE Transactions on Multimedia,* Vol. 10, No. 4, pp. 646 - 658, ISSN 1520-9210

Hsu I. C.; Chi L. P. & Bor S. S. (2009). A platform for transcoding heterogeneous markup documents using ontology-based metadata. *Journal of Network and Computer Applications,* Vol. 32, No. 3, pp. 616-629, ISSN 1084-8045

Hua Z.; Xie X.; Liu H.; Lu H. & Ma W. (2006). Design and performance studies of an adaptive scheme for serving dynamic web content in a mobile computing environment, *IEEE Transactions on Mobile Computing*, Vol. 5, No. 12, pp. -1662, ISSN 1536-1233

Huang A. W. & Sundaresan N. (2000). A semantic transcoding system to adapt web services for users with disabilities, *Proceeding of the 4th international ACM conference on Assistive technologies*, pp. 156-163, ISBN 1-58113-313-8

Hwang Y.; Kim J. & Seo E. (2003) Structure-aware web transcoding for mobile devices. *IEEE Internet Computing*, Vol. 7, No. 5, pp. 14-21, ISSN 1089-7801

Klyne G.; Reynolds F.; Woodrow C.; Ohto H.; Hjelm J.; Butler M. & Tran L. (2004). Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0, W3C, Available from http://www.w3.org/TR/CCPP-struct-vocab

Le Hégaret P.; Whitmer R. & Wood L. (2009). Document Object Model (DOM), W3C, Available from http://www.w3.org/DOM/

Lum W. Y. & Lau F. C. M. (2002). A context-aware decision engine for content adaptation. *IEEE Pervasive Computing*, Vol. 1, No. 3, pp. 41-49, ISSN 1536-1268

Manola F. & Miller E. (2004). RDF Primer, W3C, Available from http://www.w3.org/TR/rdf-primer/

McCarron S.; Ishikawa M.; Baker M.; Matsui S.; Stark P.; Wugofski T. & Yamakami T. (2010). XHTML™ Basic 1.1 - Second Edition, W3C Recommendation, Avail able from http://www.w3.org/TR/2010/REC-xhtml-basic-20101123/

Nagao K.; Shirai Y. & Squire K. (2001). Semantic annotation and transcoding: making web content more accessible, *IEEE MultiMedia*, Vol. 8, No. 2, pp. 69-81, ISSN 1070-986X

Nimmagadda Y.; Kumar K. & Lu Y. H. (2010). Adaptation of multimedia presentations for different display sizes in the presence of preferences and temporal constraints, *IEEE Transactions on Multimedia,* Vol. 12, No. 7, pp. 650 - 664, ISSN 1520-9210

Ohto H. & Hjelm J. (1999) CC/PP exchange protocol based on HTTP Extension Framework, W3C, Available from http://www.w3.org/TR/NOTE-CCPPexchange

Schubert S. & Berjon R. (2008). CSS Mobile Profile 2.0, W3C Candidate Recommendation, Available from http://www.w3.org/TR/2008/CR-css-mobile-20081210/

Wireless Application Protocol Forum, Ltd. (2001). Wireless Profiled HTTP, Available from http://read.pudn.com/downloads19/doc/comm/67224/WAP-229-HTTP-20010329-a.pdf

Wireless Application Protocol Forum, Ltd. (2001). WAG UAProf, Available from http://read.pudn.com/downloads19/doc/comm/67224/WAP-248-UAProf-20011020-a.pdf

18

Power Supply Architectures for Wireless Systems with Discontinuous Consumption

Jose Ignacio Garate and Jose Miguel de Diego
Bilbao School of Engineering-University of the Basque Country/Trelec BTC S.L.L.
Spain

1. Introduction

Many wireless systems develop in present days relay on discontinuous transmission and reception or TDD access technology, (Time Division Duplex), as this way of operation has several advantages, provided it is not required a dedicated link based only on FDD access, (Frequency Division Duplex). The used of TDD access improves the spectral efficiency, which increases the number of communication channels available sharing the same spectrum resources, besides, it enhances the energetic efficiency of the system as the hardware only works during a limited period of time and, in spite of the FDD access technique, which requires a continuous power transmission, is becoming more popular through communication systems like WIFI, MIMO and the like, the road maps for 4G radio access technologies are based on both TDD and FDD (E. Dahlman et al.,2006). The new LTE, (3GPP Long Term Evolution), standard for 3GPP flexible spectrum usage is supported through FDD/TDD harmonisation, where there is a convergence between paired spectrum and unpaired spectrum solutions (K. Fazel & S. Kaiser, 2008). To illustrate the basic time domain behaviour of the TDD and FDD channels, Fig. 1 represents their simplified time versus frequency channel distribution.
The Fig. 1 provides hints of how the communication access scheme affects several parameters of the wireless system. The present chapter discusses those issues concerning the power supply system which reverts back to the general performance of wireless systems.

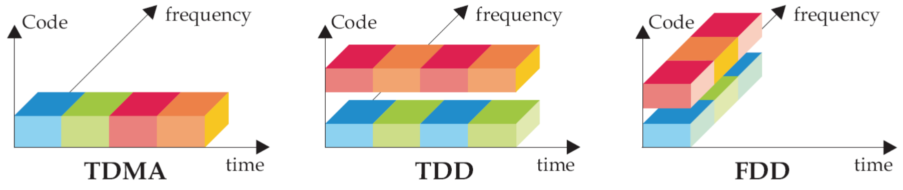


Fig. 1. TDMA, TDD and FDD time versus frequency and code channel distribution

In spite of there are many architectures of power supplies to cope with discontinuous consumption, due to the special characteristics of the wireless systems, not all of them are suitable or appropriate to all types of wireless systems. This fact is especially true or significant for battery powered wireless devices, as the restrictions in size and autonomy imposed are key factors or introduce new variables or constrains.

The research work that will be presented in this chapter is devoted to developing generic architectures of power supply systems for wireless systems, which possess the current consumption pattern of a discontinuous load. It also tries to answer, or at least eases to understand and face the design, development and production challenges related with the performance of wireless devices whenever they face with this type of current consumption.

2. Discontinuous consumption in wireless systems

As the discontinuous consumption concept is a generic topic, it requires a reference frame linked with wireless systems. This chapter considers two types of discontinuous consumption in wireless devices; a random one not directly involved in the communication process, for example, the activation of the backlights, the speaker, servos and the like, and a periodic one that will be addressed as discontinuous which is the subject of the research.
This periodic consumption is linked with the access technology employ in the wireless system and leads to the transmission and reception time periods. In spite of such classification, it is interesting to highlight that almost all tasks performed by a wireless systems processor are controlled and previously programmed, therefore, the magnitude of the current consumption, demanded by a particular event, it is predefined.

2.1 Characteristics

From the power supply perspective, one of the main attributes of a wireless system with TDD access scheme is its periodic consumption pattern, Fig. 2. The characteristics parameters of the consumption are represented in the picture and are the following:

- Period and duty cycle of the consumption (t_1, t_2).
- Magnitude of the consumption (I_{PEAK}, I_{LOAD}, $I_{STANDBY}$).
- Time mask and slopes of the communication burst.

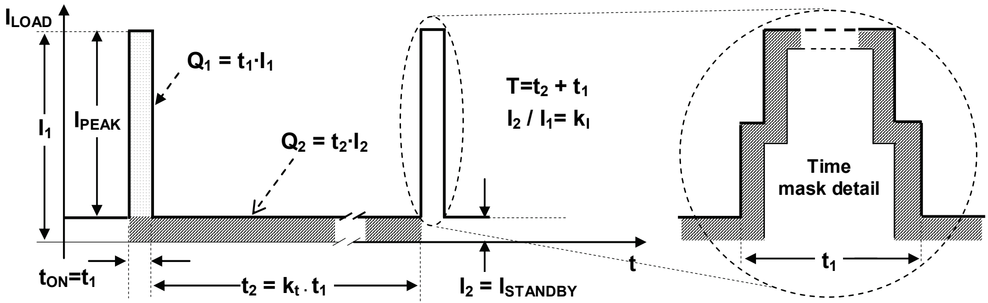


Fig. 2. Power versus time load current consumption for wireless system with discontinuous transmission, and detail of the current pulse

These parameters are the tools to determine or dimension the power supply system of a wireless system. Period, duty cycle and magnitude set the energy demands place upon the power supply. Meanwhile, the time mask and slopes of the communication burst are relevant to control the switching harmonics of the signal and, at the same time, maintain the signal spectrum within its assigned bandwidth. Fast transitions mean switching harmonics of high frequency difficult to be restrained within regulation specifications, particularly at extreme conditions of temperature and voltage.

2.2 Effects

The noticeable effects of discontinuous consumption in wireless systems are fluctuations and drops in the supply voltage, applied to the terminals of the load, around the nominal value; this fluctuation follows the consumption pattern. Voltage drop is ruled by the Ohm law, but not only must be considered the distributed resistive component of electric path between load and source, but also its reactive part. The resistive component conditions or determines the magnitude of voltage drop, meanwhile; the reactive one defines the shape and damping of consumption rise and fall slopes.

2.2.1 Voltage ripple

In wireless systems, the direct outcomes of voltage ripple are two; switching harmonics, and voltage level out of operational ranges.

A) Switching harmonics

The frequency bandwidth available for a wireless system is a scarce resource and must be optimized to allocate as many communication channels as possible. The TDD strategy to achieve this goal is multiplex in time a number of channels at the same frequency within a specific bandwidth. To make the communication systems work it is required that the transmission is produced in a specific timing. Transceiver activation, on its assigned time slot, is not produced instantaneously, which implies, before the information is received or transmitted, that there are two periods of time for conditioning the signal. These two time periods constitute the rise and fall ramp time. To this extent there are two situations to be considered:

- If ramps are too fast implies high-frequency interferences, switching harmonics. Switching harmonics reduce the amount of channel spectral density energy available for communication, consequently, they degrade the link traffic capacity and its overall performance, in other words, it means that could be set less communication links.
- If slopes are too slow, they widen the bandwidth and corrupt the spectral modulation mask, which occupy the adjacent channel reducing the traffic maximum rate and the sensitivity of adjacent receivers as their SINAD, (signal to noise ratio), is diminish.

B) Voltage ripple

The voltage level apply to the load varies between two values that correspond to minimum a maximum load. It is likely that the voltage operative range of the wireless device is exceeded in certain situations, particularly at extreme conditions of temperature.

Moreover, whenever wireless systems are battery powered, voltage drift increases as the power source voltage varies, between maximum and minimum load, due to the battery internal resistance. This is also applicable, to a certain extent, if a converter is placed between the power source and the load, as voltage drift could set the converter out of its regulation input voltage range.

2.2.2 Discontinuous current and electromagnetic compatibility

Seemingly, discontinuous consumption and voltage drops imply that the current is also variable. On the other hand, the discontinuous current drain from the power source has a direct impact on it, particularly for battery powered devices, which means energy losses in the internal battery resistance that are not uniform, as the load impedance presented varies following the consumption pattern. Besides, existence of discontinuous current implies current flux through a wire, which induces magnetic fields on the power lines.

There are three basic mechanisms or arrangements that produce magnetic fields; a signal track with a variable current, a current loop, and two parallel lines. The strength of magnetic fields varies with the level of current consumption, and their effects increase if there is any current loop involving the power lines that connect the source and load. These loops may produce interferences in any element of the wireless system, within or close to them. To make the phenomena challenging, usually, the frequency of magnetic field is a low-frequency one.

It is known that a drawback of low frequency magnetic fields is their mechanism of attenuation. Magnetic fields require an absorptive shield, (ferrite), instead of the reflective one use for high frequency electric fields, which reduces its capability to shield them. Consequently, existence of magnetic fields implies side effects, in terms of the electromagnetic compatibility, EMC, of wireless systems, which should be avoided to fulfil the applicable regulation. Thus, design requires not only a careful routing and layout of power lines but also conditions the distribution of the wireless system architecture on PCB (M. I. Montrose, 1996).

3. Power supplies and discrete components for wireless systems

From the power supply perspective, once is stated that the classification of wireless systems starts with the type of access technology employ, which also defines if the consumption is continuous or periodic, for the power supply is the subject of this chapter, wireless systems will be sorted in two generic groups based on the type of power source they employ, in spite of inherit characteristics of portable wireless systems, like cellular terminals, impose certain restrictions over the power supply architecture and the devices it made of.

3.1 Types of power sources

Power sources are sensitive to the consumption patterns of wireless systems, but the power source itself conditions the architecture of both wireless device and power supply. Consequently, wireless systems are sorted in two groups; the first are systems directly connected to the power source, and the second is made of those that require a conditioning of the power source voltage and current.

3.1.1 Direct connection to power source

Apparently, the ideal scenario may be a power supply directly connected to the wireless systems or the load. As there is no electronic between source and load, the energy losses are reduced to those in the electric paths. This is true meanwhile the energy that the load drains from the battery is constant and correctly dimensioned to its internal resistance. This ideal situation is not such, as the energy drain is not always constant, the battery discharges over time and its capacity varies over the whole operational temperature range.

Battery powered electronic devices such cellular terminals, PDAs, Ebook readers and the like are typical examples of wireless systems directly connected to the power source.

3.1.2 Voltage and current adapter

If the voltage and current levels of the source need to be conditioning, it is required a voltage converter between source and load. It does not matter if the power source is a solar panel, a battery or the mains AC power lines, this fact will only affect the architecture of the voltage converter. There are tree generic alternatives: AC-DC isolated converter, DC-DC isolated converter and DC-DC converter (B. Sahu & G.A. Rincon-mora, 2004).

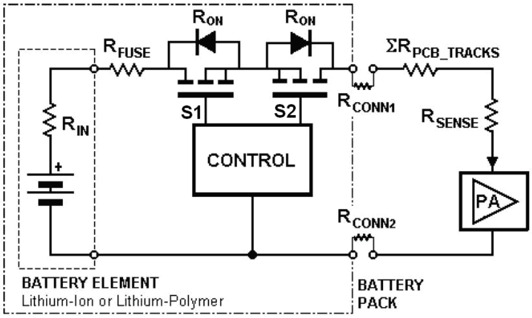
Whenever AC power source is used, it is mandatory an AC-DC isolated converter, but the need of isolation between DC power source and the wireless systems is only a matter of electromagnetic compatibility standards, electrostatic discharges and security regulation.

3.2 Systems, component and devices for wireless power supply

Unless there is a wide range of components for power supplies and sources, the next lines summarize the requirements upon key components and devices of the power supply.

3.2.1 Battery

The main power source of portable or battery powered wireless systems is the battery cell itself (Saft, 2008). The battery could be primary or secondary, i.e., rechargeable or not rechargeable, respectively. From the point of view o the chapter, the battery equivalent circuit is made of its internal resistance, R_{IN}. It use to be of low value and depends on the technology, tenths of milliohms for 1 Ahour capacity Ion-Lithium battery.



Protectionn Circuit Logic States

Voltage level	S1	S2
Low (overdischarge)	OFF	ON
Nominal	ON	ON
High (overload)	ON	OFF

Fig. 3. Detail of an Ion-Lithium battery internal protection circuit and its true table

Due to the characteristics of wireless systems stress onto battery voltage supply level, size and weight the battery technologies more suitable are, among others, the following:

- Niquel-Metal-Hydrite (NiMH) and Niquel-Cadmium (NiCd), both require fuse for safety.
- Ion-Lithium and Ion-Lithium-Polymer, both need a protection circuit plus the fuse.

The basic circuit architecture of a Lithium battery is shown in the following picture, Fig. 3. The schematic shows that the equivalent resistance of the cell is made of the internal resistance of the battery, plus the resistance of contact and the resistance of the protection circuit. The protection circuit is made of the resistance of the fuse, recommended a polyswitch type, and a couple of mosfets. The contribution of all these electronic elements must be considered as they increase the ripple of the voltage supply.

3.2.2 Converters for wireless systems. Types of converters

The performance of wireless systems is sensitive to the power supply voltage ripple and its fluctuation between maximum and minimum values. Consequently, it is highly

recommended suppress or attenuate the voltage ripple with filtering and voltage regulation. Filtering is achieved by means of high-value capacitors of low ESR and inductors; meanwhile, regulation is obtained through DC-DC converters, linear or witched ones.
As long as it is not always feasible a direct connection to the power source, power converters are used to adapt the power supply voltage and current level to those of the wireless systems, even if the power source is a battery. Moreover, depending on the systems architecture, may be required a second regulator to stabilize the output of the former one.
There is a wide range of power supply architectures available, switched or linear (R. W. Erikson, 1997). If AC-DC conversion is required, in spite of it is possible its integration within the wireless systems, is better employing an external one of a plug-in type. External AC-DCs are widespread as they ease the design and certification of the equipment electronics. This is true because external plug-in are already certified. Besides, in the particular case of wireless modules, their manufactures usually translate the discontinuous consumption impact to the application integrator or to the converter manufacturer. The Fig. 4 shows an example of such a problem; the manufacturer provides a small size chipset, already certified, but on its application note highlight that it requires to work a capacitor of the same size plus a voltage regulator.

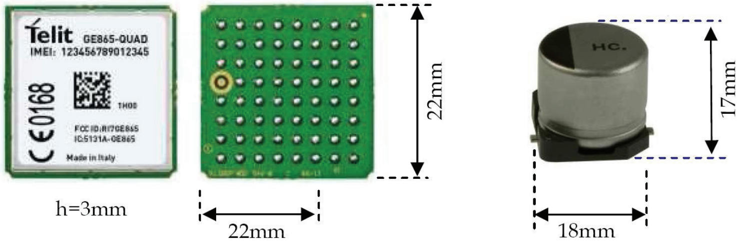


Fig. 4. Comparison between a communication module and the capacitor it requires

Summarizing the line of reasoning, the selection of power supply technologies for wireless systems should be guided by the following factors:
- Type of converter
- Isolation.
- Control scheme of the switched converter
- Control architecture of the feedback loop

Once is certified the need of power conversion, remains without answer the topic of switched or linear conversion. The advantages and drawbacks of linear regulation versus switched regulation are exposed in the following lines.
1) *Linear regulation* is obtained through a voltage control loop that samples the output voltage. The main device of a linear regulator works on its active operation region, so the voltage drop across its terminal produces power losses in the form of heat sink.
The advantages of linear regulation are its simple architecture, and the lack of electromagnetic interference. Also, it does not require inductive elements, and its current consumption under no-load conditions is low. On the other hand, it has low efficiency when the difference between input and output voltage are significant.
2) A *switched converter* employs an active device that works between cut and saturation regions; therefore, the dissipation losses are lower and cause, mainly, by switching losses

and the voltage drop in the active device over cut and saturation. The power is delivered to the load through the energy store in an inductor, which charging cycle is a function of the energy demanded by the load. So, the energy drained from the source is used mostly to feeding the load, which reduces the power losses that are limited to those of the control circuit and the component leakages. Therefore, a performance analysis of switched converters shows that they provide a better balance between input and output voltages than the linear ones. They are, also, smaller and lighter than its linear counterparts for the same power rating, mainly because the isolation transformer is smaller. Furthermore, the size and value of the transformer or the switching inductance and the capacitors are reduced as the switching frequency is increased. Lower value capacitors contribute to reduce the voltage ripple, because it is possible used ceramic capacitors of low ESR, in the order of tenths milliohms or lower.

On the other hand, a switched power supply introduces electromagnetic fields, radiated and conducted, that make the technical requirements restrictive, as the complexity of electronic design increases. Switched regulators are, also, more complex to design due to they require a higher number of discrete components, which reduces the electronic liability.

Moreover, switched converter has another issue that must be bear in mind for green design applications. As long as the current consumption is discontinuous, the load remains inactive for some periods of time; during those periods its current consumption may reach zero. Hence, switched converter has poor efficiency under no-load conditions as there is a quiescent current in the electronic of the power supply. For example, standard 12 V and 4 W commercial DC-DC have a quiescent current consumption between 30 and 50 mA.

Unless solutions switched regulation based may appear the most suitable, many manufactures employ linear regulation, especially when; there is available a power source with voltage levels close to those required by the wireless system, and size it is not a restriction. Doing so it is avoided EM fields, which increase cost and technical requirements.

3.2.3 Capacitors

Power supply of wireless systems employs capacitors to store energy and filtering. The challenges to face are finding capacitors of high value, small size and low ESR that withstand the voltage levels applied to the electronics.

Sometimes, the equipment size does not allow the use of high-value capacitors; the alternative is employ capacitors of hundred microfarads that only help to smooth voltage transitions. This is the case of GSM cellular terminals that when transmitting at maximum power, the peak current consumption may reach 3 A.

Furthermore, capacitor ESR produces load voltage ripple, and its leakage resistance introduces a continuous discharge of the battery. For example, an standard tantalum capacitor, AVX model TPCL106M006#4000, has 10 μF nominal capacitance and ESR of 4000 mΩ. An electrolytic capacitor provides higher capacitance value on a bigger size and with more ESR. On the other hand, a ceramic one has small size and low ESR, but there are not feasible for high capacitance. Table 1 highlights the differences between technologies for the same capacitance value.

Then the main limiting factors of capacitors are their ESR and size. The Table 1 provides a comparison between different types of capacitors. High value capacitors are intended to be used in the equipment, close to the load. To reduce the impact of the size it is possible; redistribute several capacitors in parallel, or use the technology of super-capacitors.

Technology	Supplier	Code	C (mF)	ESR_{MAX} (mΩ)	V_{MAX} (V)	Size (mm)
Tantalum	Kemet	A700X227M006ATE015	220	15	6,3	7.3×4.3×4.0
Tantalum	AVX	TPSD477*006-0100	470	100	6,3	7.3×4.3×2.8
Electrolytic	Nichicon	UUG1A102MNL1MS	470	790	25	Ø12.5×13.5
Tantalum	Kemet	A700X477M002ATE015	470	15	2	7.3×4.3×4.0
Tantalum	AVX	TAJD477*002-NJ	470	200	2.5	7.3×4.3×2.9
Electrolytic	Nichicon	UUG1A471MNL1MS	1000	371	10	Ø12.5×13.5
Electrolytic	Nichicon	UUG0J222MNL1MS	2200	183	6,3	Ø12.5×16.5
Electrolytic	Nichicon	UUG0J472MNL1MS	4700	100	6,3	Ø16×16.5
Electrolytic	Nichicon	UUG0J682MNL1MS	6800	77	6,3	Ø18×16.5
Super-cap	AVX	ES48301	60000	190	6,3	48×30×4.0

Table 1. Comparison between high-value capacitors technologies

Super-capacitor employs new technology developed in recent years. They combine high capacitive values with small size and low ESR, which provide good performance against high current surges, making them suitable for applications with high-peak currents. As an example, the technical parameters of some super-capacitors are summarized on Table 2.

Supplier	Code	C (mF)	ESR_{MAX} (mΩ)	V_{MAX} (V)	$I_{leakage}$ (µA max)	Size (LxWxH mm)
AVX	BZ015B603Z_B	60	96	5,5	10	28 x 17 x 6,5
AVX	BZ02CA903Z_B	90	108	12	20	48 x 30 x 6,8
Cooper	FC-3R6334-R	330	250	3,6	-	2 x 17 x 40
Maxwell	PC10-90	10	180	2,5	40	29,6x23,6x 4,8

Table 2. Comparison between super-capacitor technologies

4. Wireless systems powered through passive components

Wireless systems powered through passive components have in common the type of power source, which is often a battery. At this point, the key issue is how to increase the autonomy of these electronic devices, in doing so, the following items should be consider, balancing the tradeoffs of each one:

- Limit the load active times by reducing TX and RX periods.
- Increase the efficiency of the power supply system.
- Smooth current and voltage transitions.
- Reduce standby and quiescent current consumption.

The characteristics of battery powered wireless devices reduce the range of alternatives of power supply systems exclusively based on passive components, especially if the restrictions are combined with small size requirements. The most widespread architectures of power supply systems with passive components are described in the following topics.

Unless the conclusion and results could be extrapolated to any wireless communication system with discontinuous consumption, in order to homogenize the description, and allow the comparison of different architectures, the reference wireless communication system is a GSM cellular terminal that transmits and receives only in one time slot. In this framework, the characteristics parameters of the terminal are the following:

- Frequency of the GSM pulse = 216 Hz.
- Transmission time, t_{ON} = 1/8 of the period, or time slot that last 578 µs.
- Maximum current peak, I_{LOAD}, 2 A for a nominal 3,6 V Ion-Lithium battery.
- Standby current consumption, $I_{STANDBY}$ 20 mA @ 3,6 V.
- Mean current consumption, I_{MEAN}, equals to 2 A ·1/8 + 0,02 A ·7/8=267,5 mA @ 3,6 V, Ec. 1.

$$I_{MEAN} = I_{LOAD} \cdot \frac{t_{ON}}{T} + I_{STANDBY} \cdot \frac{(T - t_{ON})}{T} \tag{1}$$

4.1 Direct connection

Direct connection between the battery and load reduces the voltage drop in the electrical path between both elements of the systems (W. Schroeder, 2007). The Fig. 5 represents the elements that must be considered when scaling a direct connection power supply system, and it also shows the equivalent circuit of the power supply, the source and the load.

A small capacitor, C, could be included to smooth the voltage ripple of transitions between the load states ON and OFF, and it also filters some conducted emissions. For this purpose, wireless device manufactures commonly employ ceramic capacitor of around 10 µF. This capacitor only has effect on the first microseconds of the transient; consequently, the voltage drop in the load, V_{LOAD}, is the same independently of the consumption peak, Fig. 6. It could be appreciated in the figure how the voltage ripple increases proportionally to the current consumption and depends on the distributed resistance between source and load.

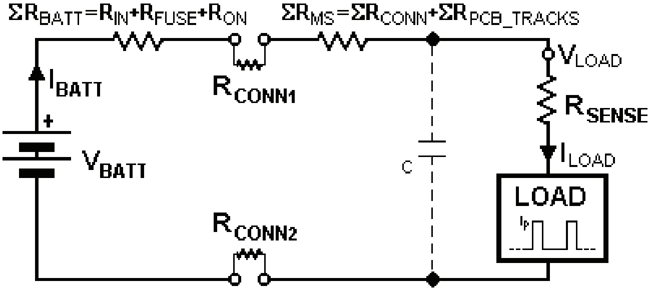


Fig. 5. Schematic diagram of a wireless system directly connected to the battery

4.2 High value load capacitor

Direct connection presents sharp transitions in the waveforms of current and voltage at both ends of load and source, Fig. 6. A straightforward regulation system uses a high-value capacitor in parallel with the load to smooth both, current and voltage, waveforms. The capacitor acts as a low-pass filter damping the slopes of the consumption transitions, in other words, it delivers a fraction of the energy that the wireless systems demands to the power source. The energy that a capacitor drives depends on its parameters, and the load consumption characteristics. Capacitor stores energy over inactive cycle of load and delivers energy when the load is active. The higher the capacitor or super-capacitor value the lower the load voltage ripple. The impact of capacitor on the power supply performance will differ depending on where is located. It could be placed in two different locations:

- In the battery cell or at the ends of the battery terminals.
- Close to the load, within the wireless electronics.

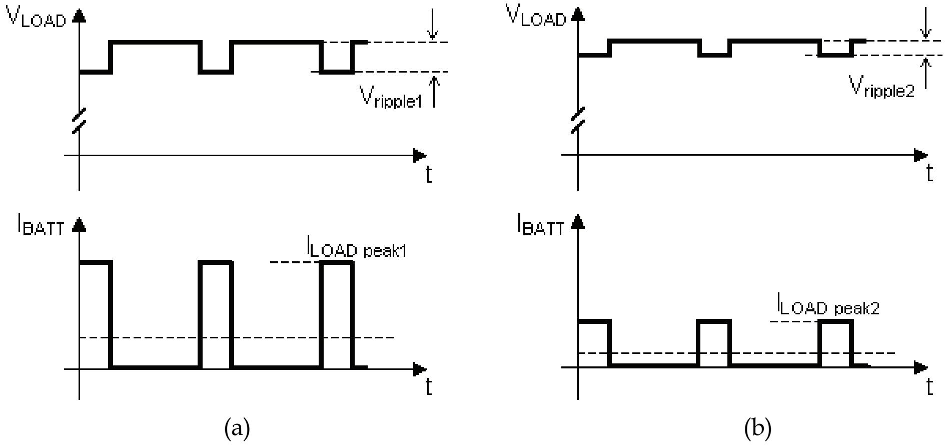


Fig. 6. Load voltage and battery current waveforms of a wireless system with discontinuous consumption for (a) maximum consumption and (b) mean consumption

Unless it may appear a satisfactory technique, it has some drawbacks. The main limiting factors of capacitors are their ESR value and size. The first produces voltage ripples, consequently. The second may lead to a capacitor size that does not fit within the wireless device. This inconvenient could be overcome, to a certain extent, by means of distribute the capacity in several capacitors in parallel or by using the technology of super-capacitors.

4.2.1 Minimum capacitor value

Before to start describing the technical alternatives of power supply systems with passive components, it is necessary made some insight concerning the minimum capacitance, C, required to absorb the current peaks at the load, which is a function of the maximum current consumption peak, its t_{ON} and the period. A straightforward way to estimate the C value is through the following reasoning line. The equivalent circuit of the power supply system plus the load, (wireless system), is presented in Fig. 7.

The circuit of the figure is valid no matter the capacitor is placed at the load or the battery, and it is made of:

- The battery of nominal voltage E.
- The distributed resistance between load and battery plus the battery internal resistance, R2.
- The ideal capacitor, C1.
- The discontinuous load made of a resistance R1 and ideal switch, S1.
- The final charge voltage, V2.
- The minimum discharge voltage, V1.
- $\Delta V=V2-V1$ is the load voltage ripple, V_{ripple}, or the magnitude of capacitive discharge.

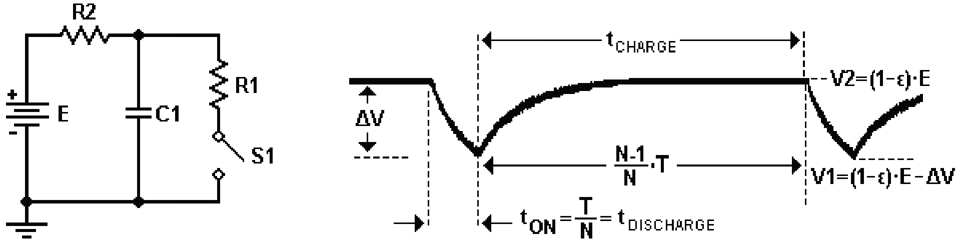


Fig. 7. Un-loaded equivalent circuit of battery, distributed resistance, capacitor and load, and detail of the load voltage ripple showing the capacitor charge and discharge

Whenever the load, or wireless system, is not activated, the capacitor voltage is equal to V1, so the capacitor discharge time is a function of $R_{LOAD}=V_{LOAD}/I_{PEAK}$, through Ec. 2.

$$t_{CHARGE} = R_2 \cdot C_1 \cdot \ln\left[\frac{E - V(0^+)}{E - V(t)}\right] \tag{2}$$

Where V1 and V2 are equal to:

$$V(0^+) = V1 = E \cdot (1-\varepsilon) - \Delta V \tag{3}$$

$$V(t) = V2 = \cdot E \cdot (1-\varepsilon) \tag{4}$$

Replacing V1 and V2, results:

$$t_{CHARGE} = R_2 \cdot C_1 \cdot \ln\left[\frac{E - E\cdot(1-\varepsilon) + \Delta V}{E - E\cdot(1-\varepsilon)}\right] = R_2 \cdot C_1 \cdot \ln\left[\frac{E\cdot\varepsilon + \Delta V}{E\cdot\varepsilon}\right] \tag{5}$$

Being the capacitor load at $V_C(0+) = V_2$, it starts a discharge cycle that last a maximum time of t_{ON}. So, the new equivalent circuit of the load plus the power supply is represented in Fig. 7. Solving the Thevenin, the circuit is simplified as it shows Fig. 8, being the Thevening voltage, E_e, equals to:

$$E_e = E \cdot \frac{1}{R2/R1 + 1} \tag{6}$$

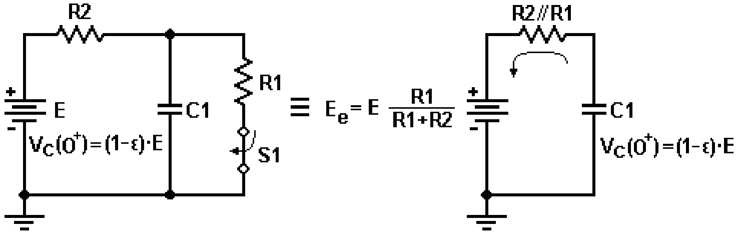


Fig. 8. Equivalent circuit of battery, distributed resistance, capacitor and load over the capacitor discharge cycle

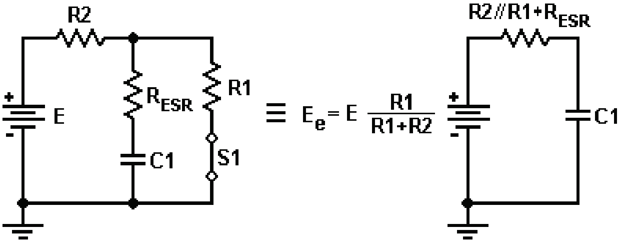


Fig. 9. Equivalent circuit of battery, distributed resistance, capacitor and load including the capacitor ESR

In these conditions, the capacitor discharge time is defined with the expression Ec. 7.

$$t_{DISCHARGE} = t_{ON} = R_1 // R_2 \cdot C_1 \cdot \ln\left[\frac{E - V(0^+)}{E - V(t)}\right] \tag{7}$$

Where V1 and V2 are equal to:

$$V(t) = V1 = \cdot E \cdot (1-\varepsilon) - \Delta V \tag{8}$$

$$V(0^+) = V2 = \cdot E \cdot (1-\varepsilon) \tag{9}$$

Replacing V1 and V2, results:

$$t_{DISCHARGE} = t_{ON} = R_1 // R_2 \cdot C_1 \cdot \ln\left[\frac{Ee - E\cdot(1-\varepsilon)}{Ee - (E\cdot(1-\varepsilon) - \Delta V)}\right] \tag{10}$$

The mathematical expressions obtained may further complicated by adding to the circuits of Fig. 7 and Fig. 8 the capacitor ESR, which is a function of the capacitance through the loss tangent, Fig. 9. The expression that relates the ESR with the capacitance is, approximately:

$$R_{ESR} = tg\delta \cdot \frac{1}{2\pi \cdot f \cdot C_1} \tag{11}$$

Consequently, the total voltage ripple, Fig. 10, is the sum of the one that causes the capacitive discharge, plus the one produce in the ESR of the capacitor is:

$$\Delta V = V_{ripple} = \Delta V_{ESR} + \Delta V_C \tag{12}$$

Bearing mind the reasoning followed on the previous lines, and replacing Ec. 5 and 12 in Ec. 5, the capacitor discharge time, with its ESR effect, is qual to:

$$t_{CHARGE} = (R_2 + R_{ESR}) \cdot C_1 \cdot \ln\left[\frac{E\cdot\varepsilon + \Delta V}{E\cdot\varepsilon}\right] \tag{13}$$

In the same way, replacing in Ec. 10, the discharge time is equal to:

$$t_{DISCHARGE} = (R_{ESR} + R_1 // R_2) \cdot C_1 \cdot \ln\left[\frac{Ee - E\cdot(1-\varepsilon)}{Ee - (E\cdot(1-\varepsilon) - \Delta V)}\right] \tag{14}$$

This lasts equations estimate the capacitance as a function of the targeted voltage ripple.

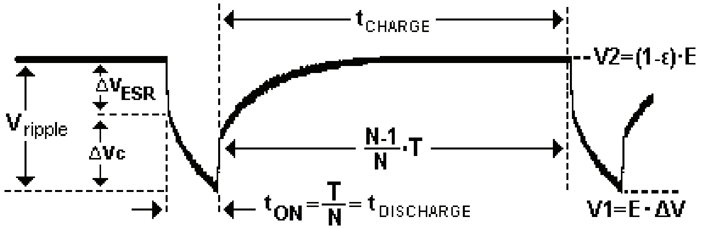


Fig. 10. Ideal waveform detail of the load voltage ripple showing the capacitor charge and discharge, and including the capacitor ESR contribution

4.2.1 At the battery ends

The first place to locate a high-value capacitor is in the battery pack. Fig. 11 shows two wireless control applications that use a high-value capacitor at the battery terminals. In (a) the maximum value is limited by the size of the mechanic, it employs two aluminium organic capacitors of 470 μF in parallel. Meanwhile in (b) the size of the equipment allows the use of a 33 mF super-capacitor.

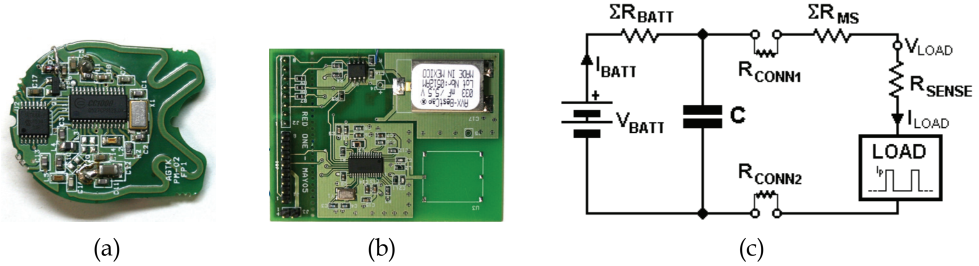


Fig. 11. Pictures of wireless control systems with capacitor place at the battery terminals, (a) high-value aluminium organic and (b) super-capacitor. (c) Schematic diagram of power supply system with high-value capacitor at the battery ends

The Fig. 11(c) shows the power supply schematic of a wireless system directly connected to the battery with a capacitor at the battery ends. The equivalent circuit is made of the resistive elements of the PCB tracks, connectors, and the equivalent resistance of the battery, which includes internal resistance, fuse resistance and protection electronics if required.

The waveforms of current and voltage at the ends of the battery represented in the Fig. 12 illustrate the behaviour of this architecture for three capacitors, it could be seen the following; the current drain from the battery, I_{BATT}, is lower than the load current demand, I_{LOAD}, the voltage ripple, V_{BATT} at the battery is lower than the ripple at the load, and, at the instant of battery connexion, the current through the connector is zero.

Therefore, place a super-capacitor or a high-value capacitor in the battery helps to reduce the space it occupied in the wireless device, but increases the size of the battery pack. Super-capacitors also presents manufacturing disadvantages as they are not suitable for automatic surface mount assembly, SMD, because do not withstand a standard lead-free oven soldering profile.

They also have some technical disadvantages. Whenever the load drains current, it creates a resistive path between the battery and load, which is made of the battery contact resistance, ΣR_{CONN}, sense resistance, R_{SENSE} and the distributed serial resistance of the PCB tracks, R_{PCB_TRACKS}. These increase the voltage drop at the load terminals.

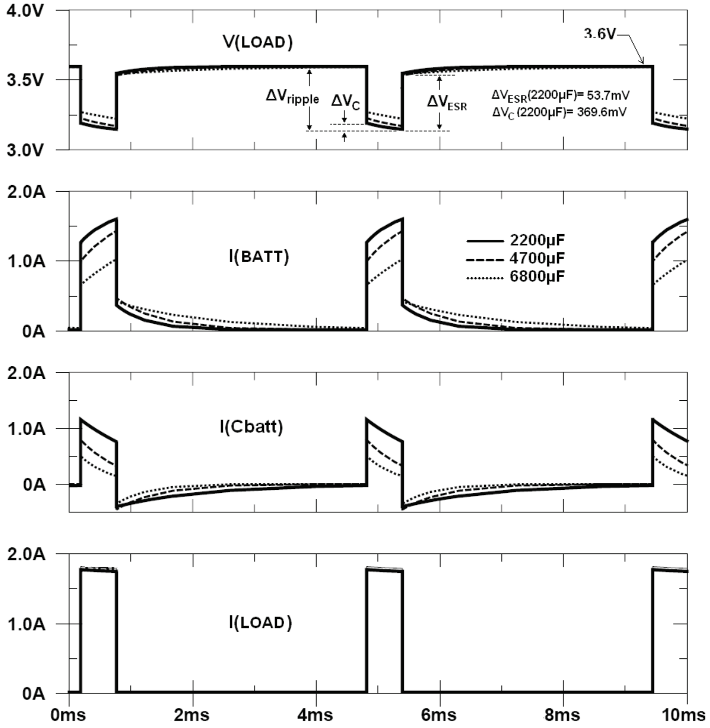
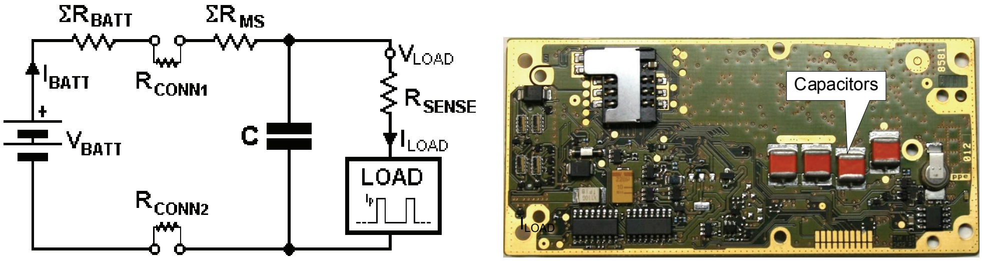


Fig. 12. Load voltage ripple, battery current, capacitor current and current load for capacitors of 2200, 4700 and 6800 µF at the battery terminals

4.2.2 At the load ends

To prevent voltage drops between the battery and load, super-capacitor or high-value capacitor should be place as close as possible to the load, as it is depicted in Fig. 13 (a). (b) is a picture of M2M wireless module with high-value capacitors at the load ends. The picture illustrates how the capacitance is distributed in several capacitors to eases fit it in the device. The total capacitance is the sum of four special tantalum capacitor of 1000 µF value each in parallel. This arrangement, not only gets high-value capacitance (4000 µF), but also reduces the equivalent ESR, as it is the sum of the ESR resistance of each capacitor in parallel.

The behaviour of this architecture is represented on Fig. 14 and Fig. 15. The first group of traces shows the input and output voltage and currents for three super-capacitors of 500, 200 and 60 mF respectively. The output voltage, V_{LOAD}, represents the magnitude of the ripple, which is a function of each capacitor ESR, as theirs ESR value is such that their charge and discharge could not be appreciated because they never fully discharge. The load current consumption, I_{LOAD}, is the result of adding battery, I_{BATT}, and capacitor, I_{Cload}, currents.



(a) (b)

Fig. 13. (a) Schematic diagram of high-value capacitor at the load terminals. (b) Picture Detail of an M2M application with tantalum capacitors at the load terminals (4000 µF)

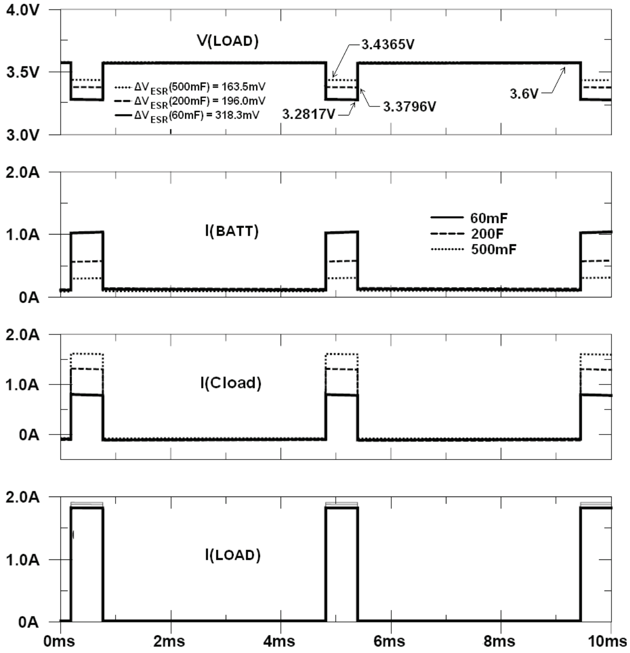


Fig. 14. Voltage ripple in the load, V(LOAD), battery, capacitor and load current with super-capacitors of 60, 200 and 500 mF

The second group reproduces the same waveforms when three electrolytic capacitors of 2200, 4700 and 6800 µF are used instead of super-capacitors. The voltage ripple is depicted as V_{LOAD} in the first trace; it shows the charge and discharge of the capacitors, and the contribution of theirs ESR to the voltage ripple.

The behaviour represented in Fig. 13 and Fig. 15 could be summarizing as follows; the current through the battery is lower than the current drain by the load, and voltage drops at the load ends are further diminished because most of the energy demanded is extracted directly from the capacitor. Unfortunately, place high-value capacitor at the load ends causes a high current peak each time the battery is replaced, Fig. 16. Eventually, this current surge may destroy or damage the connectors after a certain number of battery replacements.

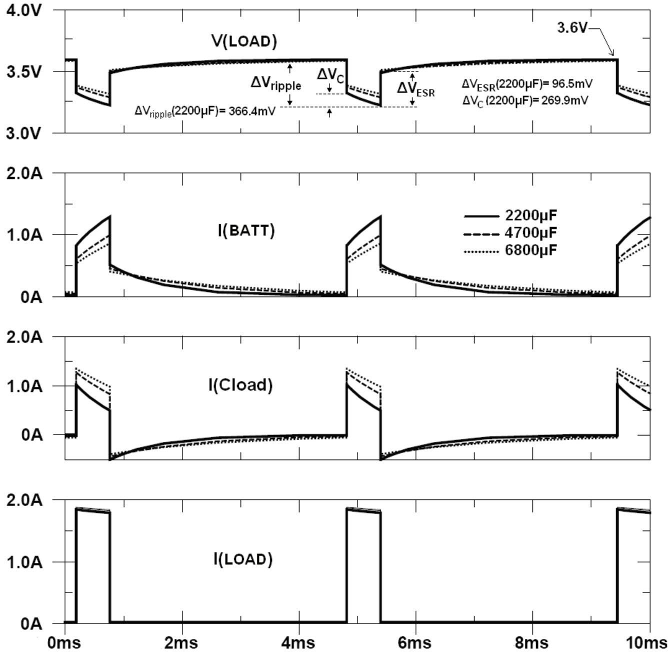


Fig. 15. (a) Voltage ripple in the load, V(LOAD), (b) battery, (c) capacitor and (d) load current with capacitors of 2200, 4700 and 6800 µF

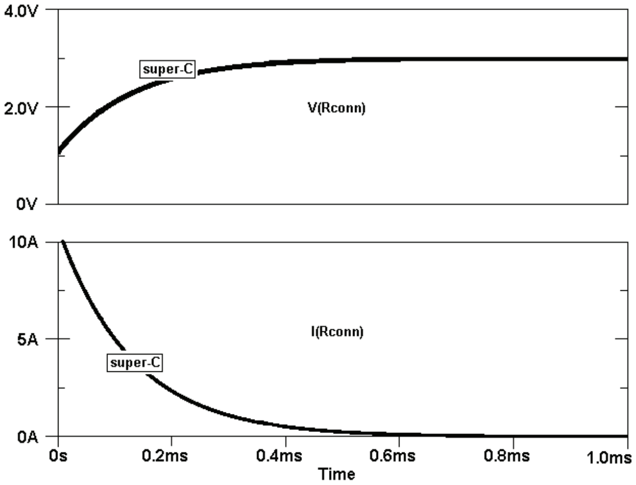
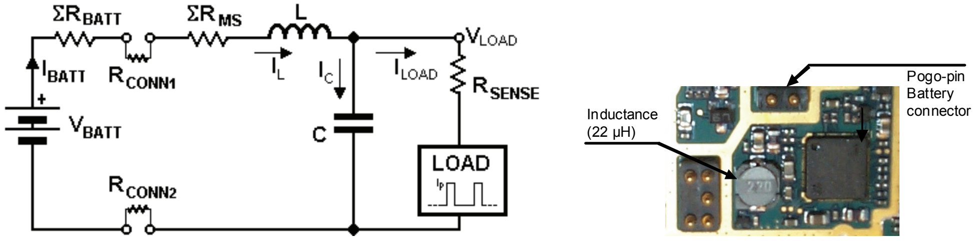


Fig. 16. Instantaneous voltage and current at the connexion of a 60 mF with the battery

4.2.3 LC network

This alternative is based on a LC network made of a series inductor and followed by a parallel capacitor, at is shown in the Fig. 17(b), and it is seldom used in older cellular terminals and radio modules with discontinuous consumption. The LC network of Fig. 17(a), from the frequency point of view, constitutes a low-pass filter, although, it also should be analyzed in the time domain to completely characterize its behaviour.

The series inductor limits the capacitor charge current; this fact smoothes the input current fluctuations, but it is required a minimum value of inductance and capacitance to be effective. The technology constrains the inductance values available through size and current parameters, which may make not feasible the required values, and consequently, the impact of an LC network is reduced to a small smooth of the transitions slopes. The table 3 illustrates SMD inductors availability of inductance higher than 680 µH that withstand currents above 2 A.



(a) (b)

Fig. 17. (a) Schematic detail of the power supply system and load with LC network.
(b) Detail of and LC network in the power supply of a cellular terminal

Core	Manufacturer	Code	L (µH)	I_{SAT}(A)	R_{DC}(mΩ)	LxWxH (mm)
Close	Bourns	SRR1240-470M	2	2	135	12,5x12,5x4,0
Close	Vishay	IHLP-4040DZ-11	0,5	0,5	270	11,3x11,5x4,0
Close	Coilcraft	MSS1246T-104	1,84	1,84	210	12,3x12,3x4,8
Open	Coilcraft	DO3340P-104	2,5	2,5	220	12,9x9,4x11,4
Open	Coilcraft	DO5040H-684	2	2	780	18,5x15,2x12,0
Open	Pulse	PF0504.104NL	2,5	2,5	153	18,5x15,2x11,4

Table 3. Comparison between inductor technologies with and without shielding

Furthermore, it must be consider that and inductor generates lines of EM fields that closes trough the air, unless the inductor posses a magnetic shield. As the EM field generated on the inductance has the same low frequency pattern of the discontinuous consumption, it implies EM interferences that could not be avoided unless shielded inductances are used. The table 3 provided also shows, for the same value and manufactures, the differences between open and shielded inductors. For example, a 100 µH shielded inductor implies a 40% increase of volume and 660 mA reduction of maximum current withstand. The space occupied by the inductor is increases by the one the capacitor requires, and in spite of capacitor may be smaller that its counterpart architectures with a single capacitor, it must be bigger that hundreds of microfarads to make the LC network effective in reducing the voltage output ripple and the input current fluctuation.

As a final remark, an LC network may generate oscillations, periodic or damped, which depend on the values of the LC network and the equivalent load resistance, as it constitutes a RLC network affected by a pulsed signal. This also means that the input current could have negative values, which implies a current send back to the battery, i.e., a charging current.

The behaviour of the LC network is represented in the following picture for three LC combinations. The first group of waveforms, Fig. 18, represents the load current, I_{LOAD}. The trace of I_{BATT} shows, how the inductance L charges over the consumption pulse, and how it delivers the energy to the capacitor over the inactive part of the cycle. At the same time, I_{LOAD} illustrates the capacitor providing current to the load meanwhile the pulse current last. When the current pulse ends the capacitor is being charged through the inductance until the next pulse came. The waveform at the capacitor ends, V_C, is equal to the load voltage, V_{LOAD}. This voltage shows the charge and discharge cycles of the capacitor which follows the current cycles of the load, I_{LOAD}. On Fig. 19 could be appreciated the same behaviour for different pairs of LC.

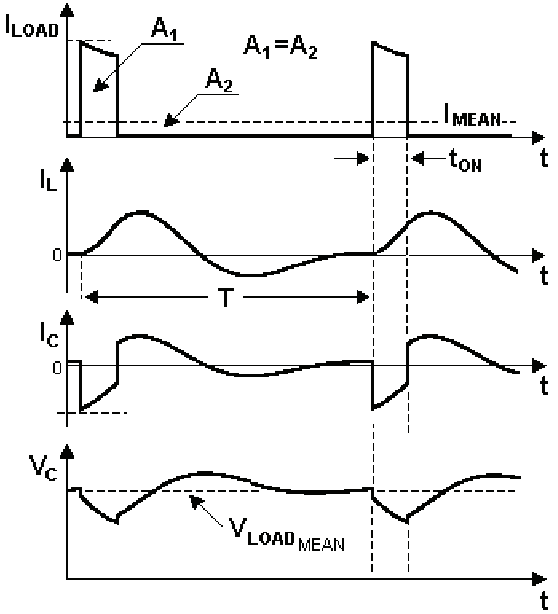


Fig. 18. Current and voltage for the power supply of a wireless system with LC filtering

5. Wireless systems powered through converters

If the power source requires a conditioning of its voltage level to those of the wireless systems internal electronics, in terms of discontinuous load, it is required a different approach to the one used for battery powered wireless systems (B. Arbetter et al., 2006). An example of this kind of wireless devices is the communication modules, called M2M (machine to machine). Whenever the power supply system employs voltage converters, no

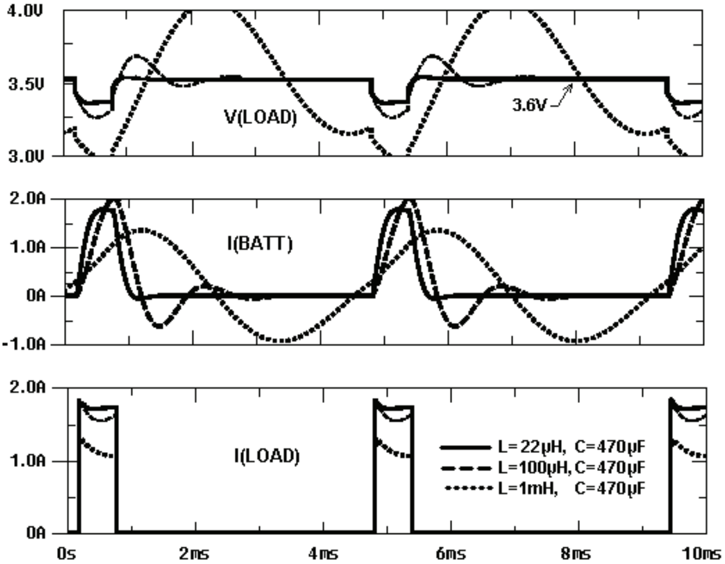


Fig. 19. Load voltage ripple, battery and load current for a capacitor of 470 μF and de 22μ, 100μ and 1mH inductors

matter if they are linear or switched, the discontinuous consumption produces current peaks at the input of the power supply. A current peak causes voltage drops in the load input voltage due to the internal resistance of the power lines and the power source connectors. If linear regulation is used, the regulators input and output current peaks are equals. On the other hand, if the converter is switched, the bigger the differences between input and output voltage, the lower the current peaks are. Nevertheless, independently of the type of power converter used, it must be design to deliver the maximum current peak.

To cope with the effects of a discontinuous consumption, in M2M modules for wireless communications, manufactures such as Wavecom, Sony Ericsson, Telit, Freescales or Siemens, recommend on theirs application notes employ high value and size capacitors, which lead to, in many design conditions, a size of recommended capacitor bigger or of the same size as the M2M module itself, at is shown in Fig. 4.

Bearing in mind what was exposed above, the technical alternatives of power supply systems with converters for wireless systems are detailed in the following headings.

5.1 Capacitor calculation

Before starting with the analysis of power supply architecture, it is required to define, on a first step, the maximum mean current that the load demands to the power source as a function of the period and duty cycle. The value obtained is used to program the current limit of the power converter. This current corresponds to the maximum current that the power converter drives to the capacitor. If it is set that the current limit must be t_{ON}/T times the maximum current peak, I_P or I_{PEAK}, that the load demands, this current limit is 1/N, (1/8 for a GSM cellular terminal). With this pattern the maximum input current of the power supply system, I_{IN}, is equal to the mean value of the maximum peak current consumption over a period. The voltage converter provides the energy that the load demands maximum consumption. Consequently, unless the maximum mean current will never overcome, I_{IN} follows the load consumption fluctuations, at it is described in the following expression:

$$I_{IN} = I_{PEAK} \cdot t_{ON} /T \tag{15}$$

Where,

- t_{ON} o $t_{discharge}$ is the duty cycle of consumption, equal to the capacitive discharge time.
- T is the consumption period.
- I_{PEAK} is the maximum peak consumption over a period.

Thus, the required capacitance is obtained through the following reasoning. Being the current through the capacitor:

$$I = C \cdot dV / dt \tag{16}$$

And considering that, on an ideal situation, the charge and discharge of a capacitor is lineal. This is feasible as the maximum drive current of the voltage converted is limited to a fixed value. Thus, doing the differential voltage equal to the voltage increment, V_{ripple}, and the differential time equal to the time increment, t_{ON}, the current is equal:

$$I = C \cdot \Delta V / \Delta t \tag{17}$$

Solving for the capacitance value, and considering that the voltage ripple ΔV_C, for an ideal capacitor, has only a capacitive discharging contribution, the capacitance results:

$$C = I \cdot t_{ON} / \Delta V_C \tag{18}$$

For example, if the mean load current is 250 mA, for a maximum peak current of 2 A, and a period time of 4,64 ms, the capacitance value, (without including the effect of its ESR), for a maximum load voltage ripple of 0,4 V, is:

$$C = 0{,}250\,\mathrm{A} \cdot 580\,\mu\mathrm{s} / 0{,}4\,\mathrm{V} \cong 3600\,\mu\mathrm{F} \tag{19}$$

5.2 Constant input current power supply

Once it is stated that a high-value capacitor smoothes current and voltage transitions and reduces its magnitude, but does not maintain constant the voltage excursion around the nominal voltage values of the power source. The first improvement could be add a voltage regulator to the capacitor, Fig. 20. The power supply system of Fig. 20 is made of a lineal or switched regulator with a fixed current limit, plus a high-value capacitor close to the wireless system load.

The Fig. 20 includes the connector and power lines resistance, R_{IN}. Following this equivalent resistance it is placed the current limited voltage regulator. The current limit of the regulator must be adjusted, approximately, to the maximum peak current averaged by N for a period. For an EGSM cellular terminal, this current is equal to the number of time slots used for transmission, $I_{PEAK}/8$. With this power supply architecture the input current, I_{IN}, always has a value close to the average current consumption. At any time, the regulator is able to provide the maximum power that the wireless system load may demand. For example, on a M2M GSM module, the input current, I_{IN}, varies following the consumption fluctuations, and never overcomes the maximum average consumption, $I_{PEAK}/8$. The R_{SENSE} resistance is used to measure the current that the capacitor drains or supplies, and, at the same time, limits the maximum current that the converter could provide. If the wireless systems requires electronics to monitoring o control the TX power it could be done by means of a current sensor, R_{SENSE}, this element increase the voltage drop and must be consider.

Unless linear conversion is an option, a switched regulator is a better solution, meanwhile the electromagnetic fields generated are under control. Linear regulation reduces the power efficiency of the supply system as the input voltage of the power source may vary over a wide range of values. As a design rule, whenever the magnitude difference between input and output voltage are not relevant a linear regulator could be used.

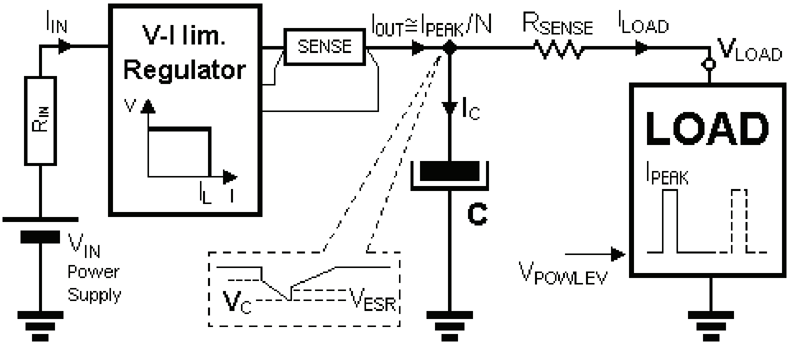


Fig. 20. Block diagram power supply system with regulated input and output capacitor

The high-value capacitor is placed at the output of the voltage converter. The capacitor stores the energy that requires the discontinuous load of the wireless system over its active time, for GSM this time corresponds to a time slot. The capacitor charge and discharge produces voltage ripple, V_C, at its terminals. Fig. 21(a) shows this voltage ripple and the regulator input current, when the load current peak is maximum, i.e., when the wireless system transmits at maximum power, and if the capacitor charge and discharge is produced at constant current.

To ease the computation of the voltage ripple an ideal capacitor is used, and two ideal sources of charge and discharge, that represents the converter and the load, respectively. Thus, the voltage ripple is a function of the current peak and capacitor value. If the capacitor is not ideal, the capacitor ESR introduces an extra voltage drop that must be added to the total voltage ripple, as it could be seen in the V_C detail of Fig. 10.

The expression for a generic wireless system with discontinuous consumption could be written as states Ec. 20, being T the consumption period, and Δt or t_{ON} the time the consumption last.

$$I_C = C \cdot \frac{dV_C}{dt} \quad ,si\ I_C = Cte \Rightarrow \Delta V = \frac{I_C}{C} \cdot \Delta t = \frac{I_C}{C} \cdot t_{ON} \tag{20}$$

Over the capacitor discharge the instantaneous voltage, V_C, could be represent as Ec. 21:

$$\text{Discharge: } \Delta V = \frac{I_{PEAK} - I_{IN}}{C} \cdot \Delta t = \frac{I_{PEAK} - I_{IN}}{C} \cdot t_{ON} \tag{21}$$

And the capacitor charging time, $V_C(t)$, is:

$$\text{Charge: } \Delta V = \frac{I_{IN}}{C} \cdot (T - t_{ON}) \Rightarrow \Delta V = \frac{I_{IN}}{C} \cdot 7 \cdot \Delta t(GSM) \tag{22}$$

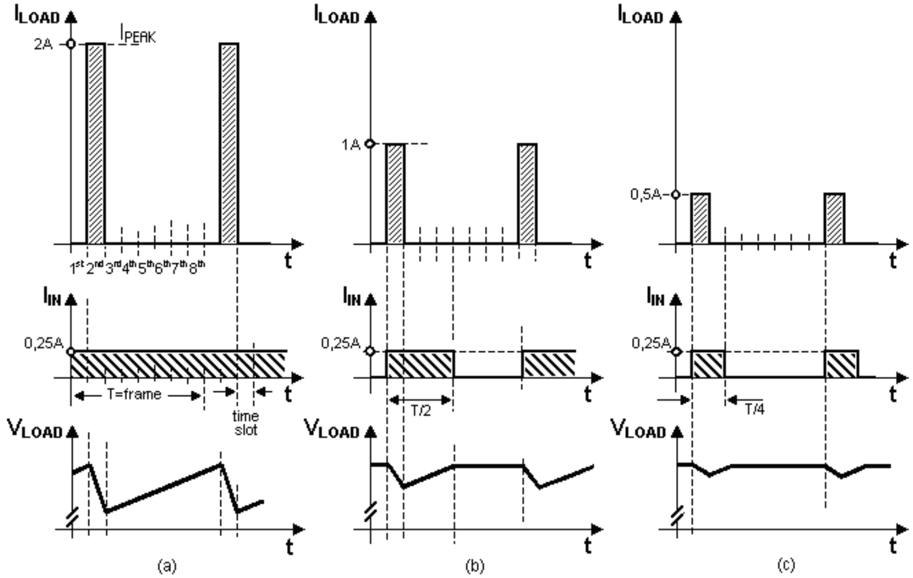


Fig. 21. PA GSM cellular terminal voltage and current waveforms for three power levels, (a) maximum and (b) (c) intermediate

At the load terminals, there is a ripple proportional to the capacitor discharge, plus the voltage drop at the current sensor, R_{SENSE}. If the sense resistance is small enough, V_C is almost equals to V_{LOAD}. The Fig. 21(a) shows the shape of the voltage ripple, V_{LOAD}, at the load ends, for different TX output power levels or discontinuous consumption.

The circuit of Fig. 20 allows to reduce the current peaks at the wireless systems input to the average value, I_{IN}, as it could be seen in Fig. 21(a). Considering the two current situations represented in the figure it could be deduce that:

- When the load is maximum, (PA transmits at its maximum power), the current consumption through the input is constant, I_{IN}, Fig. 21(a).
- If the power transmission is reduced, so the current consumption, and the current through the output, I_{IN}, is discontinuous but its maximum peak value never overcome the average current consumption for the maximum power, Fig. 21(b) and (c).

Consequently, on a GSM terminal the input current never reach a value greater than 1/8, (t_{ON}/T for a generic application), of the current peak demanded by the wireless system at maximum consumption rating.

5.3 Constant current and constant input current power supply

The Constant Input Current power supply architecture has voltage ripple at the load which varies as a function of the wireless device current consumption, Fig. 20. Unless the load voltage ripple is lower than the achieved with direct connection, this ripple could be reducing further by means of the architecture depicted in Fig. 22. This figure represents a Constant Current and Constant Input Current power supply system.

The architecture of Fig. 22 stabilizes the capacitor voltage, V_C, by means of a second regulator. This second converter could be lineal because as its input voltage range is within the same range as the output voltage, so its dropout or voltage drop in the active component is low and will not affect the efficiency of the overall power systems. The purpose of the output converter is suppress the capacitor voltage ripple. Fig. 23 shows the load voltage, V_{LOAD}, and its drive current, I_{LOAD}, obtained when using this power supply architecture. The second voltage conversion element absorbs the voltage fluctuation of charge and discharge of the capacitor in the power supply voltage, allowing a capacitor of lower value.

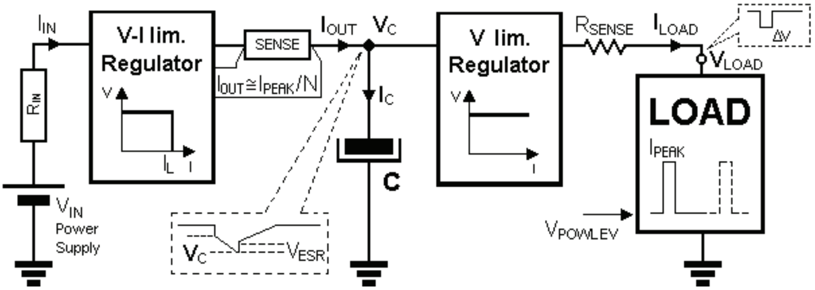


Fig. 22. Block diagram of constant current and constant input current power supply

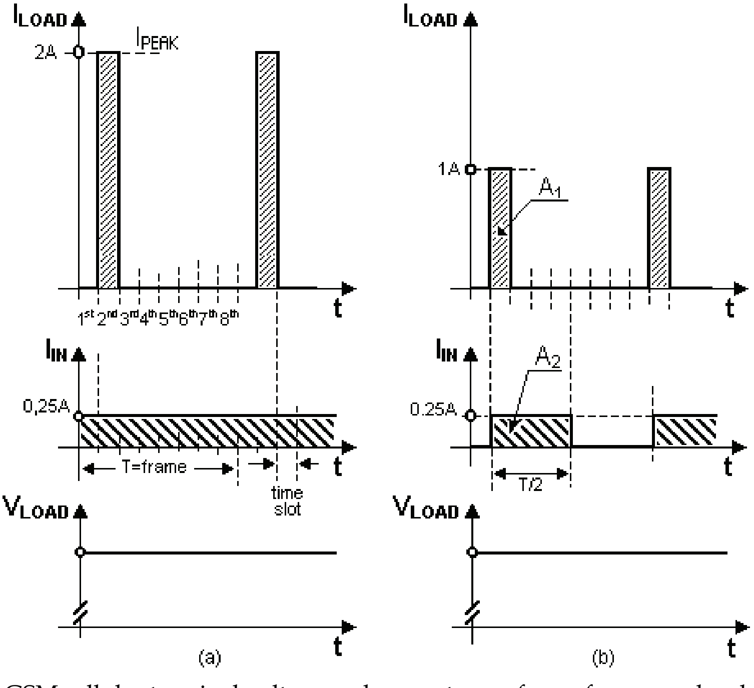


Fig. 23. PA GSM cellular terminal voltage and current waveforms for power levels,
(a) maximum and (b) intermediate

If the ripple requirements for input current and load voltage are quite restrictive, this alternative reduces its effects and, if there is also current and efficiency restrictions the linear, converter could be replace by a standard switched one of any manufacturer, for example, a MAX1678. This device has some advantage as it is specially design for GSM and UMTS, with the remark that is only efficiency in boost mode.

5.4 Summary

The ideas exposed are summarized in the Fig. 24 and Table 5. Fig. 24 provides a graphic comparison between input and output voltages and currents for the different architectures of power supply to cope with wireless systems discontinuous consumption.

The first group of waveforms, Fig. 24(a), represent the most unfavourable conditions, where the power source and the wireless systems are directly connected. The input and output current and voltages exhibit ripple with abrupt slopes.

Fig. 24(b) shows the alternative of a high value capacitor place in the load. Its charge and discharge smoothes the slopes of input and output current and voltages.

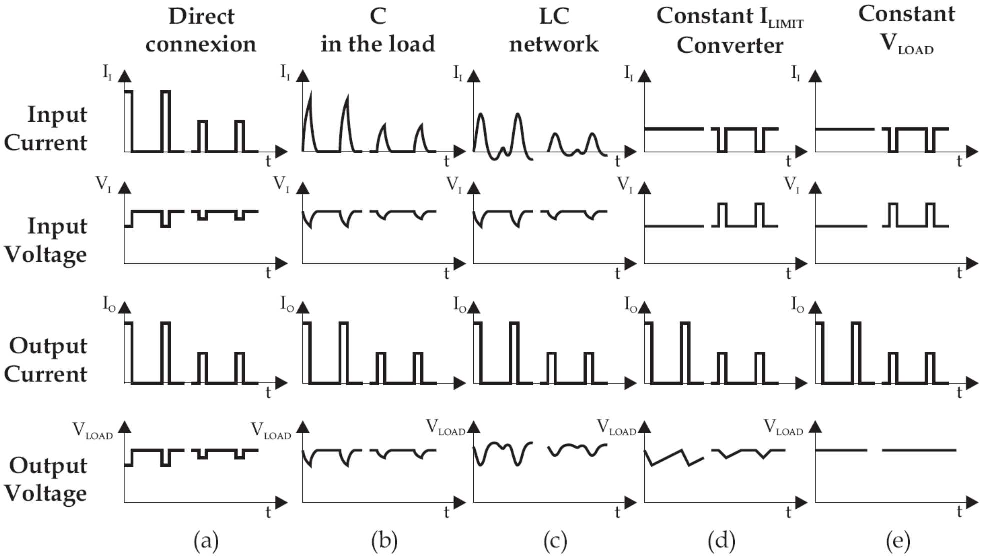


Fig. 24. Comparison between voltage and current waveforms of power supply architectures for wireless systems with discontinuous consumption

On Fig. 24(c), the alternative represented is an LC network; the graphics unveils that the input current could become negative over part of the consumption cycle, also presents over-damp, in voltage and current, that may produce oscillation.

The waveforms for regulated supply systems with fixed input current limit and a high-value capacitor are represented in Fig. 24(d). The fixed input current limit is the average value of the maximum peak current consumed in the load.

On Fig. 24(e), it is represented the voltage and current for a Constant Current and Constant Input Current Power Supply.

The Table 5 summarizes those parameters to balance in order to choose the more suitable power supply architecture for wireless systems. The table sorted the architectures in two columns, one for battery powered systems and the second for systems that require output voltage level conditioning. From the data shown, it could be inferred that: the preferred

	With passive components			With voltage converters		
	Direct connexion	**C in the battery**	**C in the load**	**LC Network**	**Constant Current Limit**	**Constant Output Voltage**
V_{IN} ripple	High	Medium /Low	Medium /Low	High	Medium /Low	Medium /Low
V_{OUT} ripple	High	Medium	Medium /Low	High	Medium /Low	Negligible
Overvoltage	No	No	No	Yes	No	No
Efficiency	High	High	Alto	High	Medium /Low	Medium /Low
EMC behaviour	Poor	Average	Average	Poor	Average	Average
Complexity	Low	Low	Low	Medium	Medium	Medium
Size	Small	Medium /High	Medium /High	Medium /High	Medium /High	Medium /High
I_{IN} pulses	High	Medium	Medium	High	Low	Low
I_{IN} peaks	Yes	Yes	Yes	Yes	No	Low

Table 5. Comparison between power supply architectures for wireless systems

architecture for battery powered systems is a high-value capacitor in the load terminals. If the wireless device needs voltage conversion, the recommended alternative is the Constant Current and Constant Input Current power supply system, made of a double voltage conversion and a high-value capacitor between the two regulators.

6. Conclusions

Wireless systems and communication electronics have their functionality and performance conditioning by the type of power consumption they present. This chapter highlights the effects of discontinuous consumption on wireless systems. It also provides keys and guidelines to identify the phenomena, and how they restrict wireless device functionality. The effects identified are:

- Power supply voltage drops produced by the current through the equivalent series resistance between source and load.
- Existence of variable electromagnetic fields, generated by the discontinuous current flux that affects the EMC performance of the electronics

Bearing in mind these two issues, the characterization of discontinuous consumption is made through the study of power supply systems suitable for such type of consumption. This is the reason why it has been proposed and analyzed two generic types of power supply systems for wireless systems that encompass all: systems directly battery powered, and systems that required supply voltage levels that differ from those provided by power source.

Commonly, power supply systems are dimensioned for the required output voltage and the maximum peak power consumption, which do not guaranty wireless systems proper operation whenever their consumption or presented load is discontinuous. Therefore, the effects of discontinuous current consumption and its solution are presented. The study analyses two power supply scenarios, direct connection between source and load through passive components, and voltage regulation. In such conditions, the most common architectures could be restricted to:

- High-value capacitor in the load or the source.
- Voltage converter with input current limit plus high-value capacitor al the load ends.
- Current Limit and Constant Input Current power supply system made of two converters, plus a high-value parallel capacitor between both converters.

Comparing and analyzing the different architectures studied, summarized in Table 5, could be concluded that suitable architectures are:

- A high-value capacitor in the load or power source.
- Current Limited and Constant Input Current power supply system if it is required conditioning the voltage levels.

Unless two regulators architecture may appear cumbersome, it is interesting highlight that increasing the switching frequency of DC-DC converters make feasible the use of such regulators, as high-frequency allows the reduction the size of the inductor and capacitors required.

In spite of there were identified valid architectures, the effects of the discontinuous consumption are not eliminated completely, as shows Fig. 24. None of the architectures gets an input current drain from the source constant.

If it is placed a high-value capacitor close to the load, there is a high current surge at the connexion instant between power source and wireless system. This current surge damages the connector reducing its lifetime. Moreover, high-value capacitor or super-capacitor may not be feasible, for they do not fit in the mechanics. Consequently, a trade-off between size and maximum available capacity and performance must be achieved.

If Current Limited and Constant Input Current power supply is used, the input current is discontinuous for consumptions below the maximum, Fig. 24(e). The direct consequence is variable current flux that produces electromagnetic fields.

7. References

B. Arbetter, R. Erikson & D. Maksimovic, *DC-DC converter design for battery-operated systems*, in Proceeding of IEEE Power Electronic Specialist Conference, 1995, vol. 1, pp. 103-109.

B. Sahu & G.A. Rincon-mora, *A Low Voltage, Non-Inverting, Dynamic, Synchronous Buck-Boost Converter for Portable Applications*, IEEE Transactions on Power Electronics, vol. 19, no. 2, Feb. 2004, pp. 443-452.

E. Dahlman, H. Ekström, A. Furuskär, Y. Jading, J. Karlsson, M. Lundevall & S. Parkvall, *The 3G Long-Term Evolution - Radio Interface Concepts and Performance Evaluation*, IEEE Vehicular Technology Conference (VTC) , Melbourne, Australia, May 2006.

K. Fazel & S. Kaiser, 2008, *Multi-Carrier and Spread Spectrum Systems: From OFDM and MC-CDMA to LTE and WiMAX*, 2nd Edition, John Wiley & Sons, ISBN 978-0-470-99821-2.

M. I. Montrose, 1996, *Printed circuit Board Design techniques for EMC Compliance*, Piscataway, NJ, IEEE Press.

R. W. Erikson, 1997, *Fundamentals of Power Electronics*, Chapman and Hall, 1st ed., New York.

Saft Rechargeable Battery Systems, 2008, *Rechargeable Battery Systems Handbook*, Available from: http://www.saftbatteries.com.

W. Schroeder, July 2007, *Direct Battery Connection Benefits Portable Designs*, White Paper, Semtech.

19

Wireless Sensor Networks in Smart Structural Technologies

Yang Wang[1] and Kincho H. Law[2]
[1]*Georgia Institute of Technology, Atlanta, Georgia,*
[2]*Stanford University, Stanford, California,*
USA

1. Introduction

Recent advances in wireless communication, as well as embedded computing, have opened many new exciting opportunities for wireless sensor networks. Miniature and low-cost wireless sensors are expected to become available in the next decade, offering countless possibilities for a wide range of applications. Among them is smart structural technology, an active research domain that holds significant promise for enhancing infrastructure management and safety. A smart structure refers to a specially equipped structure (e.g. buildings, bridges, dams, etc.) that can monitor and react to surrounding environment and the structure's own conditions, in a pre-designed and beneficial manner.

Smart structural technology encompasses at least two major fields, i.e. structural health monitoring and structural control. A structural health monitoring (SHM) system measures structural responses and predicts, identifies, and locates the onset of structural damage, e.g. due to deterioration or hazardous events. Structural sensors, such as micro-electro-mechanical system (MEMS) accelerometers, metal foil strain gages, fiber optic strain sensors, among others, have been developed and employed to collect important information about civil structures that could be used to infer the safety conditions of the structure (Farrar, *et al.* 2003, Sohn, *et al.* 2003, Chang 2009). On the other hand, structural control technology aims to mitigate adverse effects due to excessive dynamic loads (Yao 1972, Soong 1990, Housner, *et al.* 1997, Spencer and Nagarajaiah 2003).

Structural monitoring and control both involve acquiring response data in real time. In order to transmit real-time data, coaxial cables are normally employed as the primary communication link. Cable installation is labor intensive and time consuming, and can cost as much as $5,000 US dollars per communication channel (Çelebi 2002). To eradicate the high cost incurred by the use of cables, wireless systems could serve as a viable alternative (Straser and Kiremidjian 1998). Wireless communication standards, such as Bluetooth (IEEE 802.15.1), Zigbee (IEEE 802.15.4), Wi-Fi (IEEE 802.11b), are now mature and reliable technologies widely adopted in many industrial applications (Cooklev 2004). Potential applications of wireless technologies in structural health monitoring have been explored by a number of researchers, as reviewed by Lynch and Loh (2006). By incorporating a control interface, wireless sensors have also been extended to potentially command control devices for structural control applications (Wang, *et al.* 2007b).

Compared to cable-based systems, wireless structural monitoring and control systems have a unique set of advantages and technical challenges. Besides the desire for portable long-lasting energy sources, such as batteries, reliable data communication is a key issue for implementation. The purpose of this chapter is to review the important issues and metrics for adopting wireless sensor networks in smart structural systems. In a structural health monitoring system, sensors are typically deployed in a passive manner, primarily for measuring structural responses. Structural control systems, on the other hand, need to respond in real time to mitigate excess dynamic response of structures. Typical feedback control systems require real-time information and measurements to instantly determine control decisions. Although structural monitoring and control applications pose different needs and requirements, efficient information flow plays a key and critical role in both implementations. For example, the transmission latency and limited bandwidth of wireless devices can impede real-time operations as required by control or monitoring systems. In addition, communication in a wireless network is inherently less reliable than that in cable-based systems, particularly when node-to-node communication range lengthens. These information constraints, including bandwidth, latency, range, and reliability, need to be considered carefully using an integrated system approach and pose many challenges in the selection of hardware technologies and the design of software/algorithmic strategies.

The chapter adopts a previously designed wireless structural monitoring and control system as an example to discuss various intriguing research challenges (Wang, *et al.* 2005, Wang 2007). The system contains wireless sensing and control units that can be used for both wireless structural health monitoring and real-time feedback structural control. Modularized software is designed for the wireless units, so that application programs can be conveniently embedded into the units. The architectural details of the wireless structural monitoring and control system are presented. For different structural applications, including health monitoring and control, special communication protocols have been designed to efficiently manage the information flow among the wireless units. Finally, laboratory and field validation tests have been conducted to assess the performance of the prototype wireless structural monitoring and control system.

2. Design and implementation of a wireless sensing and control unit

Sensing and control units are the fundamental components of a wireless monitoring and control system. The prototype wireless unit is designed in such a way that the unit can serve as either a sensing unit (i.e. a unit that collects data from sensors and wirelessly transmits the data), a control unit (i.e. a unit that calculates optimal control decisions and commands control devices), or a unit for both sensing and control. Fig. 1 shows the functional diagram of the prototype wireless sensing and control unit. The wireless sensing unit shown in the top part of Fig. 1 serves as the core component, with which off-board modules for signal conditioning and signal generation can be easily incorporated.

2.1 Hardware and software of the wireless sensing and control unit

The wireless sensing unit consists of three functional modules: sensor signal digitization, computational core, and wireless communication. The sensing interface converts analog sensor signals into digital data, which is then transferred to the computational core through a high-speed Serial Peripheral Interface (SPI) port. Besides a low-power 8-bit Atmel ATmega128 microcontroller, external Static Random Access Memory (SRAM) is integrated

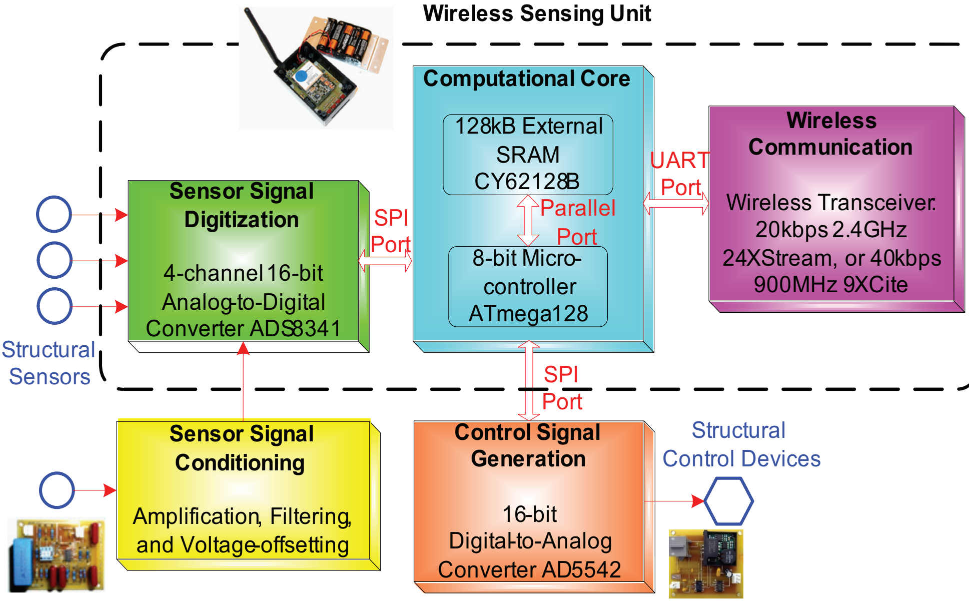

Wireless Sensing Unit
Computational Core
128kB External SRAM CY62128B
Parallel Port
8-bit Micro-controller ATmega128
UART Port
Wireless Communication
Wireless Transceiver: 20kbps 2.4GHz 24XStream, or 40kbps 900MHz 9XCite
Sensor Signal Digitization
4-channel 16-bit Analog-to-Digital Converter ADS8341
SPI Port
Structural Sensors
SPI Port
Sensor Signal Conditioning
Amplification, Filtering, and Voltage-offsetting
Control Signal Generation
16-bit Digital-to-Analog Converter AD5542
Structural Control Devices

Fig. 1. Functional diagram detailing the hardware design of the wireless sensing unit. Additional off-board modules can be interfaced to the wireless sensing unit to condition sensor signals and issue control commands

with the computational core to accommodate local data storage and analysis. The computational core communicates with a wireless transceiver (24XStream or 9XCite models currently provided by Digi International) through a Universal Asynchronous Receiver and Transmitter (UART) interface. The auxiliary sensor signal conditioning module assists in amplifying, filtering, and offsetting analog sensor signals prior to digitization. The auxiliary control signal generation module offers an interface through which the wireless sensor can send analog control commands to structural control devices. Hardware design of the wireless unit and auxiliary modules have been described in details elsewhere (Wang, *et al.* 2005, Wang 2007, Wang, *et al.* 2007a). The key parameters of the prototype wireless sensing unit are summarized in Table 1. Peer-to-peer communication among wireless units is supported for collaborative data analysis.

In order to manage the hardware components in a wireless sensing unit, software modules are implemented and embedded in the ATmega128 microcontroller. For the ATmega128 microcontroller, software can be written in a high-level programming language, such as C, compiled into binary instructions, and loaded into the non-volatile flash memory of the microcontroller. When the wireless unit is powered on for normal operation, the microcontroller automatically starts executing the embedded instructions. The software design of the wireless sensing and control units follows the hierarchical structure as shown in Fig. 2. At the bottom level are the software modules that manage the basic peripherals of the microcontroller. The middle layer consists of software modules that manage other onboard hardware components. Specific software modules for structural health monitoring and control are implemented in the top level application layer.

Design Parameter	**Specification**	
Computing Core		
Microcontroller	8-bit RISC[1] architecture, up to 16MIPS[2] throughput at 16MHz	
Flash Memory	128K bytes	
Internal SRAM[3]	4K bytes	
External SRAM	128K bytes	
EEPROM[4]	4K bytes	
Power Consumption	30mA active, 55μA standby	
Wireless Transmission	*9XCite*	*24XStream*
Operating Frequency	ISM 902-928 MHz	ISM 2.4000 - 2.4835 GHz
Data Transfer Rate	38.4 kbps	19.2 kbps
Communication Range	Up to 300' (90m) indoor, 1000' (300m) at line-of-sight	Up to 600' (180m) indoor, 3 miles (5km) at line-of-sight
Power Consumption	55mA transmitting, 35mA receiving, 20μA standby	150mA transmitting, 80mA receiving, 26μA standby
Sensing Interface	4 channels, 16-bit, up to 100 kHz	
Control Interface	1 channel, 16-bit, up to 1 MHz	
Physical Size	10.2cm × 6.5cm × 4.0cm	

Table 1. Key parameters of the wireless sensing unit

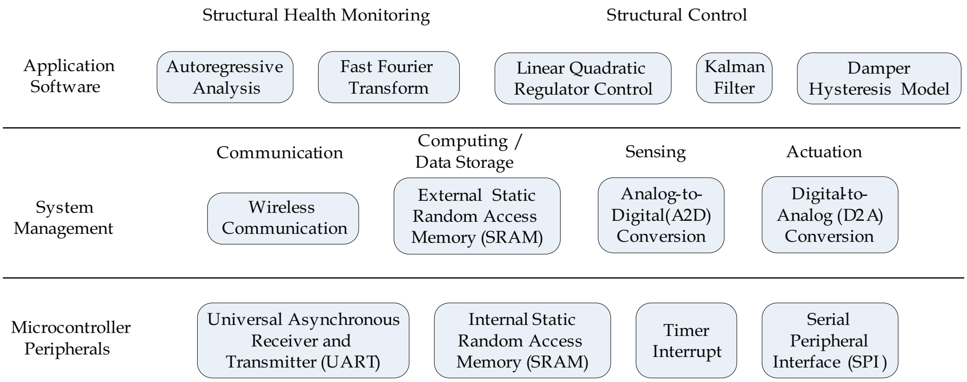

Structural Health Monitoring
Structural Control
Application Software
Autoregressive Analysis
Fast Fourier Transform
Linear Quadratic Regulator Control
Kalman Filter
Damper Hysteresis Model
Communication
Computing / Data Storage
Sensing
Actuation
System Management
Wireless Communication
External Static Random Access Memory (SRAM)
Analog-to-Digital(A2D) Conversion
Digital-to-Analog (D2A) Conversion
Microcontroller Peripherals
Universal Asynchronous Receiver and Transmitter (UART)
Internal Static Random Access Memory (SRAM)
Timer Interrupt
Serial Peripheral Interface (SPI)

Fig. 2. Three-layer software architecture for the ATmega128 microcontroller in the wireless sensing and control unit

[1] RISC: reduced instruction set computer.

[2] MIPS: million instructions per second.

[3] SRAM: static random access memory.

[4] EEPROM: electrically erasable programmable read-only memory.

As shown in Fig. 2, the lowest level of the embedded software manages the peripherals of the ATmega128 microcontroller and serves as the fundamental modules to support the functions of other hardware components. Embedded modules include: timer interrupt functions, byte-by-byte communication through the UART and SPI ports, and internal memory management. The timer interrupt service is implemented to achieve a constant time step for sensor data sampling. The interrupt function is also a powerful feature that allows the software to momentarily pause an executing task (such as data processing or wireless communication) to sample data from the sensing interface according to a precise timing schedule. Immediately after servicing the sensing interface, the paused task is resumed and the program continues its execution. This timer interrupt feature is utilized to implement continuous data streaming from multiple wireless sensing units, where sensor data sampling has to occur at a constant sampling step amidst the execution of the wireless communication or data interrogation program. In effect, the software supports concurrency thereby allowing multiple software tasks to execute at the same time.

Building on top of the microcontroller peripherals are the software drivers that manage other hardware components in the wireless unit. Utilizing the UART peripheral, the wireless communication driver provides the following functions interfacing the microcontroller with the wireless transceiver: 1) reading or setting the radio parameters of the attached wireless transceiver; 2) sending or receiving data through the wireless transceiver; 3) implementing the state machine representing the wireless communication protocol. A driver module is implemented to manage the 128kB external Static Random Access Memory (SRAM). This module includes functions to enable and disable the external SRAM, as well as functions that allow access to the lower 64kB half or higher 64kB half of the memory chip. The other two hardware drivers, the A2D and the D2A modules, manage the interfaces with the structural sensors and control devices. The ATmega128 microcontroller provides only one SPI port, which is shared by both the A2D converter (ADS8341) for sensing and the D2A converter (AD5542) for control. The A2D module commands the ADS8341 to convert a 0 to 5V analog sensor signal into a 16-bit integer. Knowing the sensitivity and offset of the sensor signal, the microcontroller can then compute a floating-point number quantifying the physical parameter being measured by the sensor. Conversely, the D2A module takes a floating-point number between -5V and 5V as input, converts the number into a 16-bit integer, and pushes the integer to the AD5542 to output the corresponding control voltage signal.

Utilizing the hardware drivers for communication, computing, sensing, and control, software can be developed to support structural health monitoring and control applications. A number of engineering algorithms, such as Fast Fourier Transform (FFT), autoregressive (AR) analysis, linear quadratic regulator (LQR) control, and Kalman Filter, have been implemented and embedded in the wireless units. The ability to execute embedded application software allows the wireless sensing units to make and execute decisions. Onboard data processing also helps save energy resources (i.e. preserving limited battery power) by reducing wireless transmission of large amounts of raw sensor data. With the application software executing in the wireless unit, each unit acts as an autonomous node in a wireless monitoring and control network. This architecture of distributed sensing and control represents a new paradigm in structural health monitoring and control, as opposed to traditional centralized systems, where data are processed in a centralized location.

2.2 Communication constraints

As noted in Table 1, the sensing unit is designed to support two wireless transceivers: 900-MHz 9XCite and 2.4-GHz 24XStream (MaxStream 2004, MaxStream 2005). This dual transceiver support allows the wireless sensing and actuation unit to operate in different regions around the world. Wireless communication poses four major constraints to the information flow within a structural monitoring and control network: bandwidth, latency, reliability, and range. It is thus important to assess the communication constraints of the transceivers.

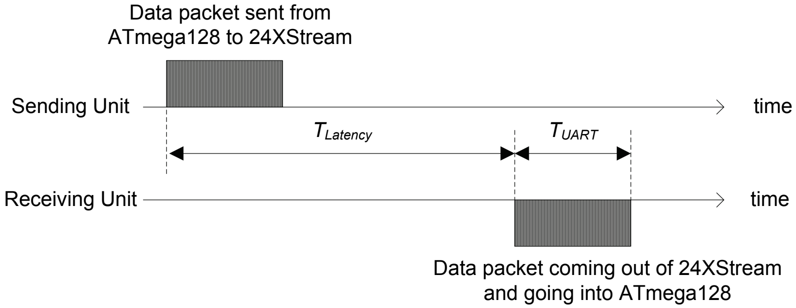

Data packet sent from ATmega128 to 24XStream
Sending Unit
time
$T_{Latency}$
T_{UART}
Receiving Unit
time
Data packet coming out of 24XStream and going into ATmega128

Fig. 3. Three-layer software architecture for the ATmega128 microcontroller in the wireless sensing and control unit

Bandwidth and latency are about the timing characteristics of the communication links. Bandwidth refers to the data transfer rate once a communication link is established. Using the MaxStream 24XStream transceiver as an example, the anticipated transmission time for a single data packet is illustrated in Fig. 3. The transmission time consists of the communication latency, $T_{Latency}$, of the transceivers and the time to transfer data between the microcontroller and the transceiver using the universal asynchronous receiver and transmitter (UART) interface, T_{UART}. Assume that the data packet to be transmitted contains N bytes and the UART data rate is T_{UART} bps (bits per second), which is equivalent to R_{UART} /10 bytes per second, or R_{UART} /10000 bytes per millisecond. It should be noted that the UART is set to transmit 10 bits for every one byte (8 bits) of sensor data, including one start bit and one stop bit. The communication latency in a single transmission of this data packet can be estimated as:

$$T_{SingleTransm} = T_{Latency} + \frac{10000N}{R_{UART}} \text{ (ms)} \tag{1}$$

In the prototype wireless sensing and control system, the setup parameters of the 24XStream transceiver are first tuned to minimize the transmission latency, $T_{Latency}$. Then experiments are conducted to measure the actual achieved $T_{Latency}$, which turns out to be around 15±0.5ms. The UART data rate of the 24XStream radio, R_{UART}, is selected as 38400 bps in the implementation. For example, if a data packet sent from a sensing unit to a control unit contains 11 bytes, the total time delay for a single transmission is estimated to be:

$$T_{SingleTransm} = 15 + \frac{10000 \times 11}{38400} \approx 17.86 \text{ (ms)} \tag{2}$$

This amount of latency typically has minimal effect in most monitoring applications, but has noticeable effects to the timing-critical feedback control applications. This single-transmission delay represents one communication constraint that needs to be considered when calculating the upper bound for the maximum sampling rate of the control system. A few milliseconds of safety cushion time at each sampling step are a prudent addition that allows a certain amount of randomness in the wireless transmission latency without undermining the reliability of the communication system. Although the achievable transmission latency, $T_{Latency}$, is around 15ms for the MaxStream 24XStream transceiver, it can be as low as 5ms for the 9XCite transceiver. This lower latency makes the 9XCite transceiver more suitable for real-time feedback control applications compared with the 24XStream transceiver. However, the 9XCite transceiver may only be used in countries and regions where the 900MHz band is for free public usage, such as the North America, Israel, South Korea, among others. On the other hand, operating in the 2.4GHz international ISM (Industrial, Science, and Medical) band, the 24XStream transceiver can be used in most countries in the world.

The other two constraints, reliability and range, are related to the attenuation of the wireless signal traveling along the transmission path. The path loss *PL* (in decibel) of a wireless signal is measured as the ratio between the transmitted power, $P_{TX}[\text{mW}]$, and the received power, $P_{RX}[\text{mW}]$ (Molisch 2005):

$$PL[\text{dB}] = 10\log_{10}\frac{P_{TX}[\text{mW}]}{P_{RX}[\text{mW}]} \tag{3}$$

Path loss generally increases with the distance, *d*, between the transmitter and the receiver. However, the loss of signal strength varies with the environment along the transmission path and is difficult to quantify precisely. Experiments have shown that a simple empirical model may serve as a good estimate to the mean path loss (Rappaport and Sandhu 1994):

$$\overline{PL}(d)[\text{dB}] = PL(d_0)[\text{dB}] + 10n\log_{10}\left(d/d_0\right) + X_\sigma[\text{dB}] \tag{4}$$

Here $PL(d_0)$ is the free-space path loss at a reference point close to the signal source (d_0 is usually selected as approximately 1 meter). X_σ represents the variance of the path loss, which is a zero-mean log-normally-distributed random variable with a standard deviation of σ. The parameter *n* is the path loss exponent that describes how fast the wireless signal attenuates over distance. Basically, Eq. (4) indicates an exponential decay of signal power:

$$P_{RX}[\text{mW}] = P_0[\text{mW}]\left(d/d_0\right)^{-n} \tag{5}$$

where P_0 is the received power at the reference distance d_0. Typical values of *n* are reported to be between 2 and 6. Table 2 shows examples of measured *n* and σ values in different buildings for 914 MHz signals (Rappaport and Sandhu 1994).

A link budget analysis can be used to estimate the range of wireless communication (Molisch 2005). To achieve a reliable communication link, it is required that

$$P_{TX}[\text{dBm}]+AG[\text{dBi}] \geq PL(d)[\text{dB}] + RS[\text{dBm}] + FM[\text{dB}] \tag{6}$$

where *AG* denotes the total antenna gain for the transmitter and the receiver, *RS* the receiver sensitivity, *FM* the fading margin to ensure quality of service, and $PL(d)$ the realized path loss at some distance *d* within an operating environment. Table 3 summarizes the link budget analysis for the 9XCite and 24XStream transceivers, and their estimated indoor ranges.

Building	n	σ[dB]
Grocery store	1.8	5.2
Retail store	2.2	8.7
Suburban office building – open plan	2.4	9.6
Suburban office building – soft partitioned	2.8	14.2

Table 2. Values of path loss exponent *n* at 914MHz

	9XCite	24XStream
P_{TX} [dBm]	0.00	16.99
AG [dBi]	4.00	4.00
RS [dBm]	-104.00	-105.00
FM [dB]	22.00	22.00
$\overline{PL}$ = P_{TX} +*AG*-*RS*-*FM* [dB]	86.00	103.99
$PL(d_0)$ [dB], d_0 = 1 m	31.53	40.05
$\overline{PL} - PL(d_0)$ [dB]	54.47	63.94
n	2.80	2.80
$\overline{d}$ [m]	88.20	192.18

Table 3. Link budget analysis to the wireless transceivers

The path loss exponent *n* is selected to be 2.8, which is the same as the soft-partitioned office building in Table 2. Generally, 2.4GHz signals typically have higher attenuation than 900MHz signals, and, thus, a larger path loss exponent *n*. The transmitter power P_{TX}, receiver sensitivity *RS*, and fading margin *FM* of the two wireless transceivers are obtained from the MaxStream datasheets. A total antenna gain *AG* of 4 is employed by assuming that low-cost 2dBi whip antennas are used by both the transmitting and the receiving sides. The free-space path loss at d_0 is computed using the Friis transmission equation (Molisch 2005):

$$PL(d_0)[\text{dB}] = 20\log_{10}\left(4\pi d_0/\lambda\right) \tag{7}$$

where λ is the wavelength of the corresponding wireless signal. Finally, assuming that the variance X_σ is zero, the mean communication range $\overline{d}$ can be derived from Eq. (4) as:

$$\overline{d} = d_0 10^{\left(\overline{PL} - PL(d_0)\right)/(10n)} \tag{8}$$

Table 3 shows that the transceivers can achieve the communication ranges indicated in Table 1. It is important to note the sensitivity of the communication range with respect to the path loss exponent *n* in Eq. (8). For instance, if the exponent of 3.3 for indoor traveling (through brick walls, as reported by Janssen & Prasad (1992) for 2.4 GHz signals) is used for the 24XStream transceiver, its mean communication range reduces by half to 87m.

3. Wireless structural health monitoring

The prototype wireless unit is first investigated for applications in wireless structural health monitoring. A structural health monitoring system measures structural performance and operating conditions with various types of sensing devices, and evaluates structural safety using damage diagnosis or prognosis methods. Eliminating lengthy cables, wireless sensor networks can offer a low-cost alternative to traditional cable-based structural health monitoring systems. Another advantage of a wireless system is the ease of relocating sensors, thus providing a flexible and easily reconfigurable system architecture. This section first provides an overview to the wireless structural health monitoring system, and then introduces the communication protocol design for reliable data management in the prototype system. A large-scale field deployment of the wireless structural health monitoring system is summarized at the end of the section.

3.1 Overview of the wireless structural health monitoring system

A simple star-topology network is adopted for the prototype wireless sensing system. The system includes a server and multiple structural sensors, signal conditioning modules, and wireless sensing units (Fig. 4). The server is used to organize and collect data from multiple wireless sensing units in the sensor network. The server is responsible for: 1) commanding all the corresponding wireless sensing units to perform data collection or interrogation tasks, 2) synchronizing the internal clocks of the wireless sensing units, 3) receiving data or analysis results from the wireless network, and 4) storing the data or results. Any desktop or laptop computer connected with a compatible wireless transceiver can be used as the server. The server can also provide Internet connectivity so that sensor data or analysis results can be viewed remotely from other computers over the Internet. Since the server and the wireless sensing units must communicate frequently with each other, portions of their software are designed in tandem to allow seamless integration and coordination.

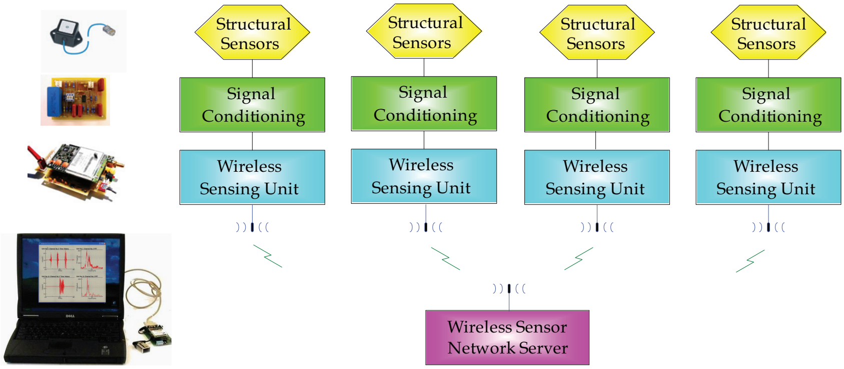


Fig. 4. An overview of the prototype wireless structural sensing system

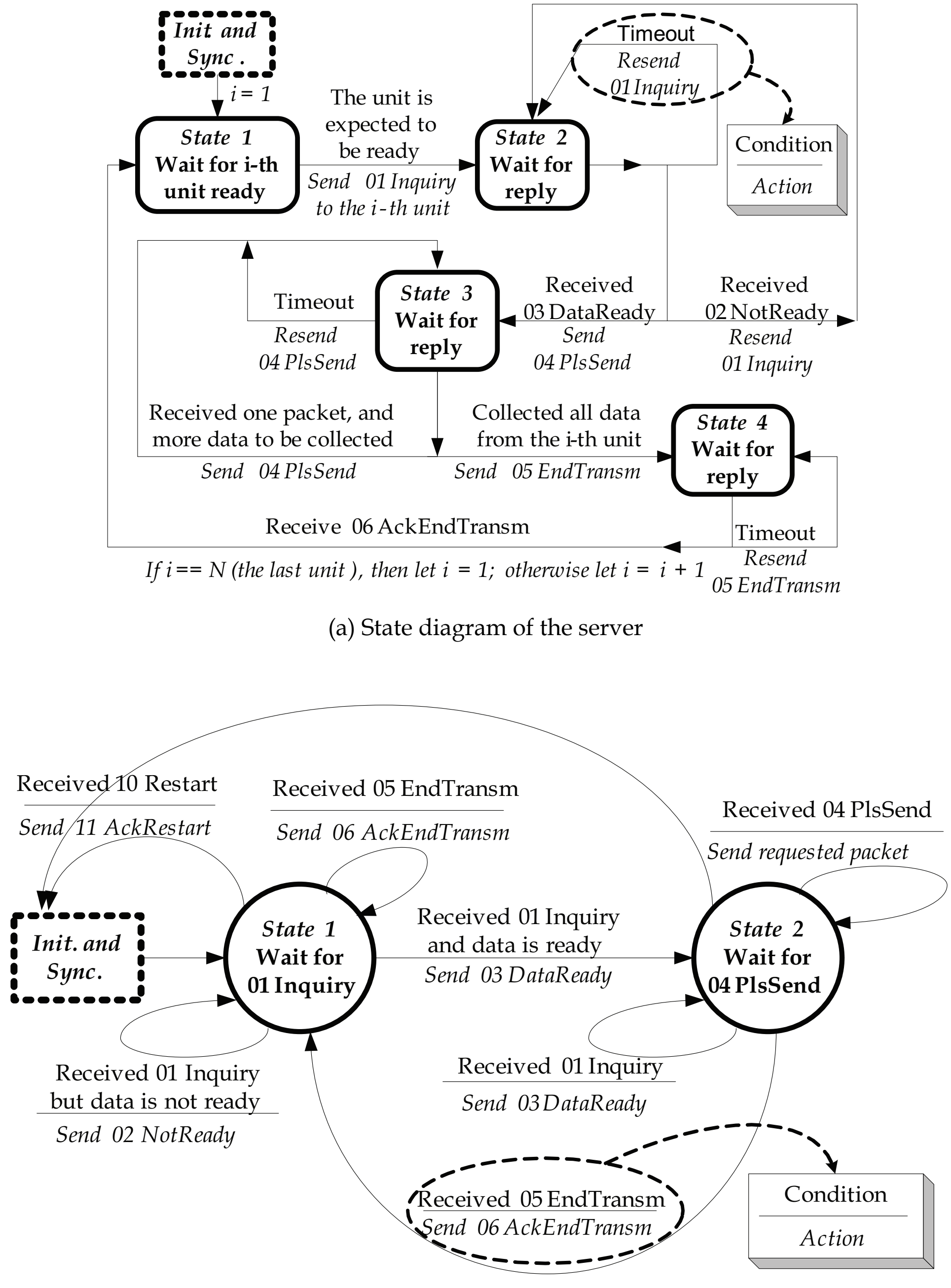
At the beginning of each wireless structural sensing operation, the server issues commands to all the units, informing the units to restart and synchronize. After the server confirms that all the wireless sensing units have restarted successfully, the server queries the units one by one for the data they have thus far collected. Before the wireless sensing unit is queried for its data, the data is temporarily stored in the unit's onboard SRAM memory buffer.

A unique feature of the embedded wireless sensing unit software is that it can continue collecting data from interfaced sensors in real-time as the wireless sensing unit is transmitting data to the server. In its current implementation, at each instant in time, the server can only communicate with one wireless sensing unit. In order to achieve real-time continuous data collection from multiple wireless sensing units with each unit having up to four analog sensors attached, a dual stack approach has been implemented to manage the SRAM memory (Wang, *et al.* 2007a). When a wireless sensing unit starts collecting data, the embedded software establishes two memory stacks dedicated to each sensing channel for storing the sensor data. For each sensing channel, at any point in time, only one of the stacks is used to store the incoming data stream. While incoming data is being stored into the dedicated memory stack, the system transfers the data in the other stack out to the server. For each sensing channel, the role of the two memory stacks alternate as soon as one stack is filled with newly collected data.

3.2 Communication design of the wireless structural health monitoring system

To ensure reliable wireless communication between the server and the wireless units, the communication protocol needs to be carefully designed and implemented. The commonly used network communication protocol is the Transmission Control Protocol (TCP) standard. TCP is a sliding window protocol that handles both timeouts and retransmissions. It establishes a full duplex virtual connection between two endpoints. Although TCP is a reliable communication protocol, it is too general and cumbersome to be employed by the low-power and low data-rate communication such as in a wireless structural sensing network. The relatively long latency of transmitting each wireless packet is another bottleneck that may slow down the communication throughput. For practical and efficient application in a wireless structural sensing network, a simpler communication protocol is needed to minimize transmission overhead. Yet the protocol has to be designed to ensure reliable wireless transmission by properly addressing possible data loss. The communication protocol designed for the prototype wireless sensing system inherits some useful features of TCP, such as data packetizing, sequence numbering, timeout checking, and retransmission. Based upon pre-assigned arrangement between the server and the wireless units, the sensor data stream is segmented into a number of packets, each containing a few hundred bytes. A sequence number is assigned to each packet so that the server can request the data sequentially.

To simplify the communication protocol, special characteristics of the structural health monitoring application are exploited. For example, since the objective in structural monitoring application is normally to transmit sensor data or analysis results to the server, the server is assigned the responsibility for ensuring reliable wireless communication. As the server program normally runs on a computer and the wireless unit program runs on a microcontroller, it is also reasonable to assign the responsibility to the server since it has much higher computing power. For example, communication is always initiated by the server. After the server sends a command to the wireless sensing unit, if the server does not receive an expected response from the unit within a certain time limit, the server will resend the last command again until the expected response is received. However, after a wireless sensing unit sends a message to the server, the unit does not check if the message has arrived at the server correctly or not, because the communication reliability is assigned to the server. The wireless sensing unit only becomes aware of the lost data when the server queries the unit for the same data again. In other words, the server plays an “active” role in the communication protocol while the wireless sensing unit plays more of a “passive” role.



(a) State diagram of the server

(b) State diagram of a wireless sensing unit.

Fig. 5. Communication state diagrams for wireless structural health monitoring

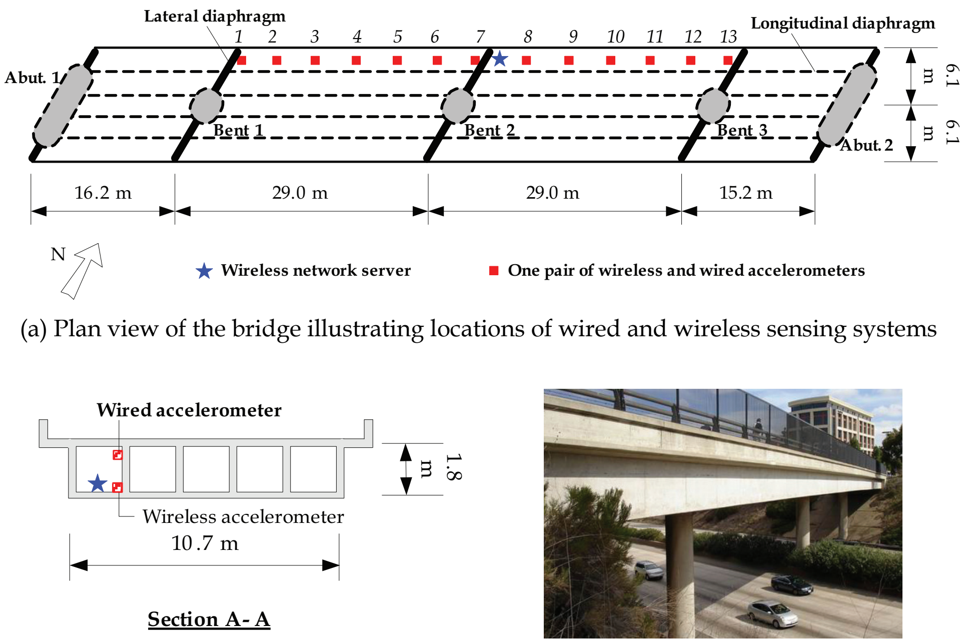
Finite state machine concepts are employed in designing the communication protocol for the wireless sensing units and the server. A finite state machine consists of a set of states and definable transitions between the states (Tweed 1994). At any point in time, the state machine can only be in one of the possible states. In response to different events, the state machine transits between its discrete states. The communication protocol for initialization and synchronization can be found in (Wang, *et al.* 2007a). Fig. 5(a) shows the communication state diagram of the server for one round of sensor data collection, and Fig. 5(b) shows the corresponding state diagram of the wireless units. During each round of data collection, the server collects sensor data from all of the wireless units; note that the server and the units have separate sets of state definitions.

At the beginning of data collection, the server and all the units are all set in State 1. Starting with the first wireless unit in the network, the server queries the sensor for the availability of data by sending the '01Inquiry' command. If the data is not ready, the unit replies '02NotReady', otherwise the unit replies '03DataReady' and transits to State 2. After the server ensures that the data from this wireless unit is ready for collection, the server transits to State 3. To request a data segment from a unit, the server sends a '04PlsSend' command that contains a packet sequence number. One round of data collection from one wireless unit is ended with a two-way handshake, where the server and the unit exchange '05EndTransm' and '06AckEndTransm' commands. The server then moves on to the next unit and continuously collects sensor data round-by-round.

3.3 Field validation tests at Voigt Bridge

Laboratory and field validation tests have been conducted to verify the performance of the wireless structural monitoring system. Field tests are particularly helpful in assessing the limitations of the system, and providing valuable experience that can lead to further improvements in the system hardware and software design. This section presents an overview of the validation tests conducted on the Voigt Bridge located on the campus of the University of California, San Diego (UCSD) in La Jolla, California (Fraser, *et al.* 2006). Voigt Bridge is a two lane concrete box girder highway bridge. The bridge is about 89.4m long and consists of four spans (Fig. 6). The bridge deck has a skew angle of 32^o, with the concrete box-girder supported by three single-column bents. Over each bent, a lateral diaphragm with a thickness of about 1.8m stiffens the girder. Longitudinally, the box girder is partitioned into five cells running the length of the bridge (Fig. 6b).

Girder cells along the north side of the bridge are accessible through four manholes on the bridge sidewalk. As a testbed project for structural health monitoring research, a cable-based system has been installed in the northern-most cells of the box girder. The cable-based system includes accelerometers, strain gages, thermocouples, and humidity sensors. For the purpose of validating the proposed wireless structural monitoring system, thirteen accelerometers interfaced to wireless sensing units are installed within the two middle spans of the bridge to measure vertical vibrations. One wireless sensing unit (associated with one signal conditioning module and one accelerometer) is placed immediately below the accelerometer associated with the permanent wired monitoring system. While the wired accelerometers are mounted to the cell walls, wireless accelerometers are simply mounted on the floor of the girder cells to expedite the installation process. The installation and calibration of the wireless monitoring system, including the placement of the 13 wireless sensors, takes about an hour. The MaxStream 9XCite wireless transceiver operating at 900MHz is integrated with each wireless sensing unit.



(a) Plan view of the bridge illustrating locations of wired and wireless sensing systems

(b) Elevation view to section A-A

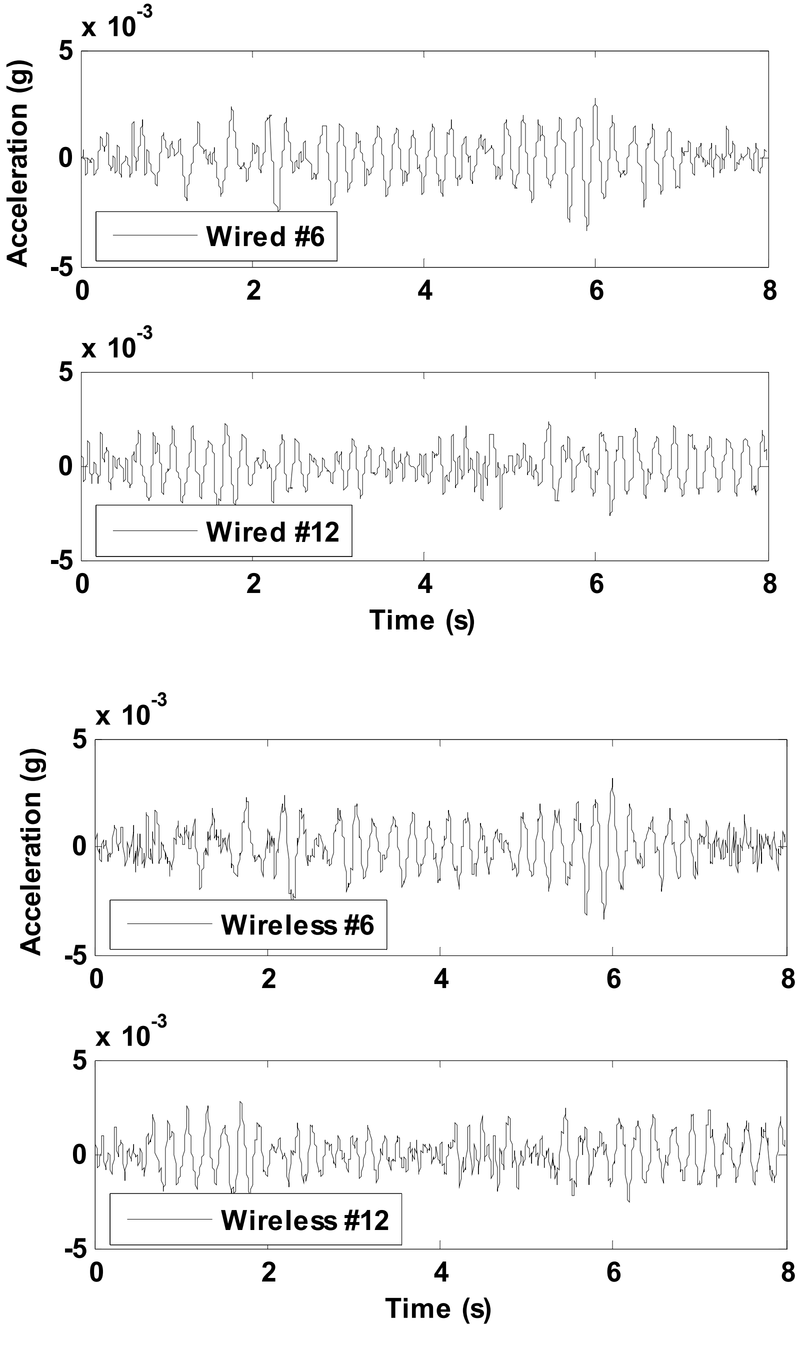
(c) Side view of the bridge over Interstate 5

Fig. 6. Voigt Bridge test comparing the wireless and wired sensing systems

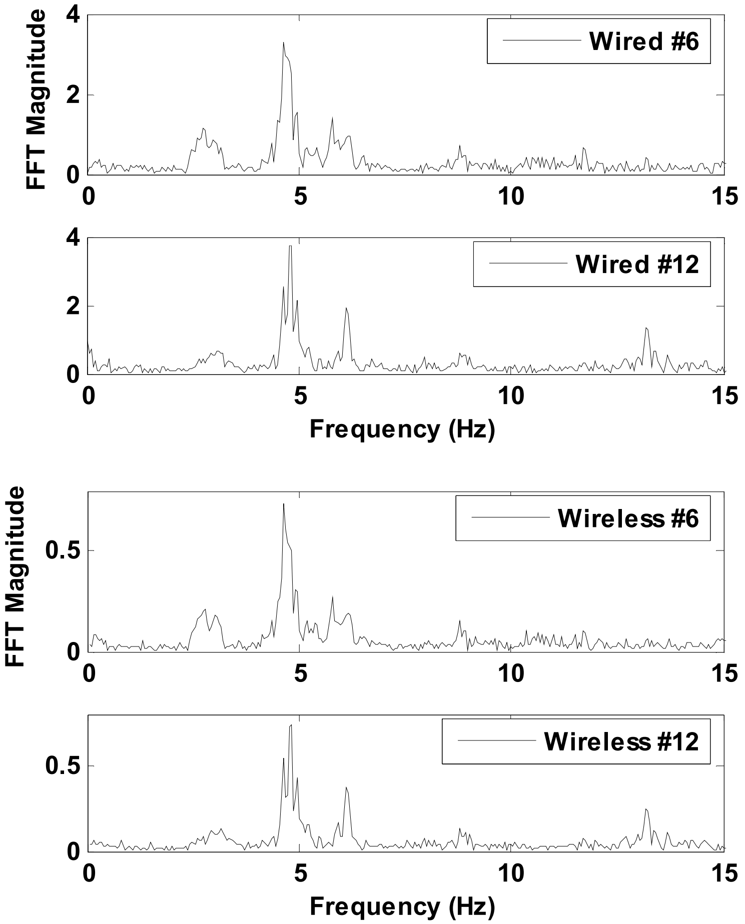
Two types of accelerometers are associated with each monitoring system. At locations #3, 4, 5, 9, 10, and 11 in Fig. 6(a), PCB Piezotronics 3801 accelerometers are used with both the cabled and the wireless systems. At the other seven locations, Crossbow CXL01LF1 accelerometers are used with the cabled system, while Crossbow CXL02LF1Z accelerometers are used with the wireless system. Table 4 summarizes the key parameters of the three types of accelerometers. Signal conditioning modules are used for filtering noise, amplifying and shifting signals for the wireless accelerometers. The signals of the wired accelerometers are directly digitized by a National Instruments PXI-6031E data acquisition board (Fraser, *et al.* 2006). Sampling frequencies for the cable-based system and the wireless system are 1,000 Hz and 200 Hz, respectively.

Specification	PCB3801	CXL01LF1	CXL02LF1Z
Sensor Type	Capacitive	Capacitive	Capacitive
Maximum Range	± 3g	± 1g	± 2g
Sensitivity	0.7 V/g	2 V/g	1 V/g
Bandwidth	80 Hz	50Hz	50Hz
RMS Resolution (Noise Floor)	0.5 mg	0.5 mg	1 mg
Minimal Excitation Voltage	5 ~ 30 VDC	5 VDC	5 VDC

Table 4. Parameters of the accelerometers used by the wire-based and wireless systems in the Voigt Bridge test



(a) Comparison between wired and wireless time history data

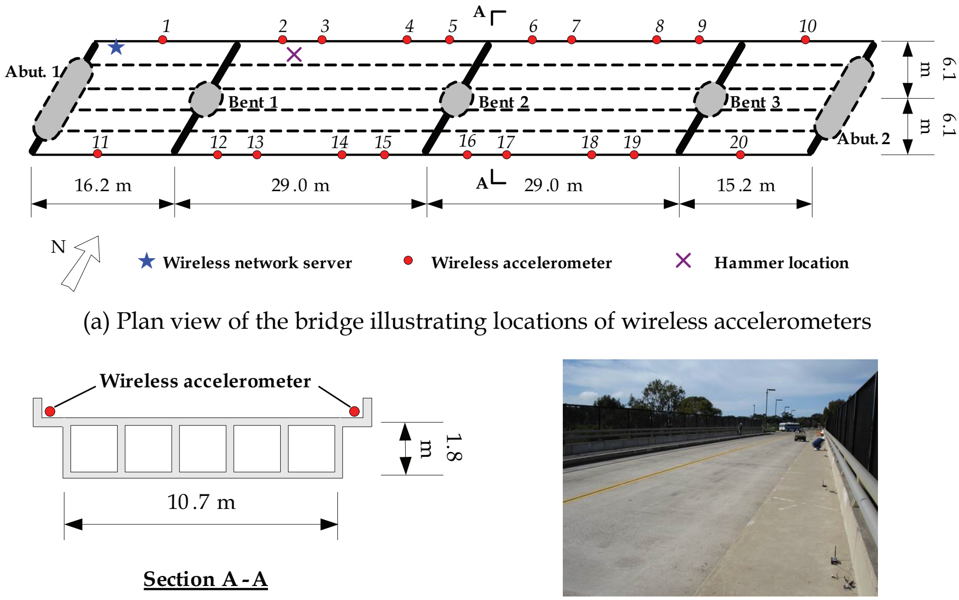


(b) Comparison between FFT to the wired data, as computed offline by a computer, and FFT to the wireless data, as computed online by the wireless sensing units

Fig. 7. Comparison between wired and wireless data for the Voigt Bridge test

The bridge is under normal traffic operation during the tests. Fig. 7(a) shows the time history data at locations #6 and #12, collected by the cable-based and wireless monitoring systems when a vehicle passes over the bridge. A close match is observed between the data collected by the two systems. The minor difference between the two data sets can be mainly attributed to two sources: 1) the signal conditioning modules are used in the wireless system but not in the cabled system; 2) the wired and wireless accelerometer locations are not exactly adjacent to each other, as previously described. Fig. 7(b) shows the Fourier spectra determined from the time history data. The FFT results using the data collected by the cabled system are computed offline, while the FFT results corresponding to the wireless data are computed online in real-time by each wireless sensing unit. After each wireless sensing unit executes its FFT algorithm, the FFT results are wirelessly transmitted to the network server. Strong agreement between the two sets of FFT results validates the computational accuracy of the wireless sensing units. It should be pointed out that because the sampling frequency of the cabled system is five times higher than that of the wireless system, the magnitude of the Fourier spectrum for the wired data is also about five times higher than those for the wireless data.

One attractive feature of the wireless sensing system is that the locations of the sensors can be re-configured easily. To determine the operating deflection shapes of the bridge deck, the configuration of the original wireless sensing system is changed to attain a more suitable spatial distribution. Twenty wireless accelerometers and the wireless network server are mounted to the bridge sidewalks (Fig. 8). The communication distance between the server and the farthest wireless sensing unit is close to the full length of the bridge. The installation and calibration of the wireless monitoring system, including the placement of all the wireless sensors, again takes about an hour. Sampling frequency for the wireless monitoring system is kept at 200 Hz.



(a) Plan view of the bridge illustrating locations of wireless accelerometers

(b) Elevation view to section A-A (c) Side view of the bridge over Interstate 5

Fig. 8. Wireless accelerometer deployment for the operating deflection shape analysis to Voigt Bridge

The communication protocol described before is implemented in the server and the wireless sensing units. For the tests described in this chapter, the server collects sensor data or FFT results from all 20 wireless units. Due to the length of the bridge and continuous traffic conditions, the wireless communication experienced some intermittent difficulty during the two days of field testing. However, the wireless monitoring system proved robust by recognizing communication failures and successfully retransmitting the lost data according to the communication protocol rules.

Fig. 9 shows the operating deflection shapes (ODS) extracted from one set of test data collected during a hammer excitation test. The hammer excitation is applied at the location shown in Fig. 8(a) and during intervals of no passing vehicles. DIAMOND, a modal analysis software package, is used to extract the operating deflection shapes (ODS) of the bridge deck (Doebling, *et al.* 1997). Under hammer excitation, the operating deflection shapes at or near a resonant frequency should be dominated by a single mode shape (Richardson 1997). Fig. 9 presents the first four dominant operating deflection shapes of the bridge deck using wireless acceleration data. The ODS #1 (4.89 Hz), #2 (6.23 Hz), and #4 (11.64 Hz) show primarily flexural bending modes of the bridge deck; a torsional mode is observed in ODS #3 (8.01 Hz). Successful extraction of the ODS shows that the acceleration data from the 20 wireless units are well synchronized.

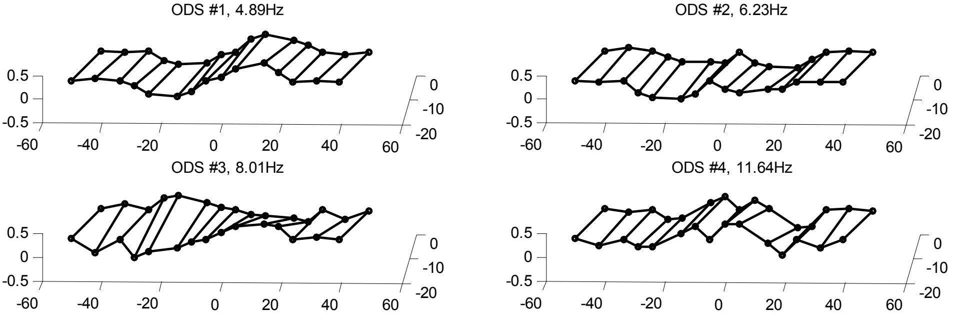


Fig. 9. Operating deflection shapes extracted from wireless sensor data

4. Wireless structural control

A feedback structural control system contains an integrated network of sensors, controller, and control devices. When external excitation (such as an earthquake or typhoon) occurs, structural response is measured by sensors and immediately collected by the controller. The controller makes optimal decisions for the control devices, which then exert appropriate forces to the structure so that undesired structural vibrations are effectively mitigated. A wireless sensing/control unit can serve as both the sensor and the controller modules of a structural control system. Each wireless unit, in addition to collecting and communicating sensor data in real time, can also make optimal control decisions and command control devices. This section first provides an overview to the prototype wireless structural control system, and then describes the communication protocol design of the system. Laboratory wireless structural control experiments are also reported.

4.1 Overview of the wireless structural control system

Fig. 10 illustrates the communication patterns of a centralized control system using cabled communication and the prototype decentralized structural control system using wireless communication. In a centralized structural control system, one centralized controller collects data from all the sensors in the whole structure, computes control decisions, and then dispatches command signals to control devices. This centralized control strategy implemented with cabled communication requires high instrumentation cost, is difficult to reconfigure, and potentially suffers from single-point failure at the controller. Wireless decentralized control architectures can offer an alternative solution. In a decentralized architecture, multiple sensors and controllers can be distributively placed in a large structure, where the controller nodes can be closely collocated with the control devices. As each controller only needs to communicate with sensors and control devices in its vicinity, the requirement on communication range can be significantly reduced, and the communication latency decreases by reducing the number of sensors or control devices that each controller has to communicate with.

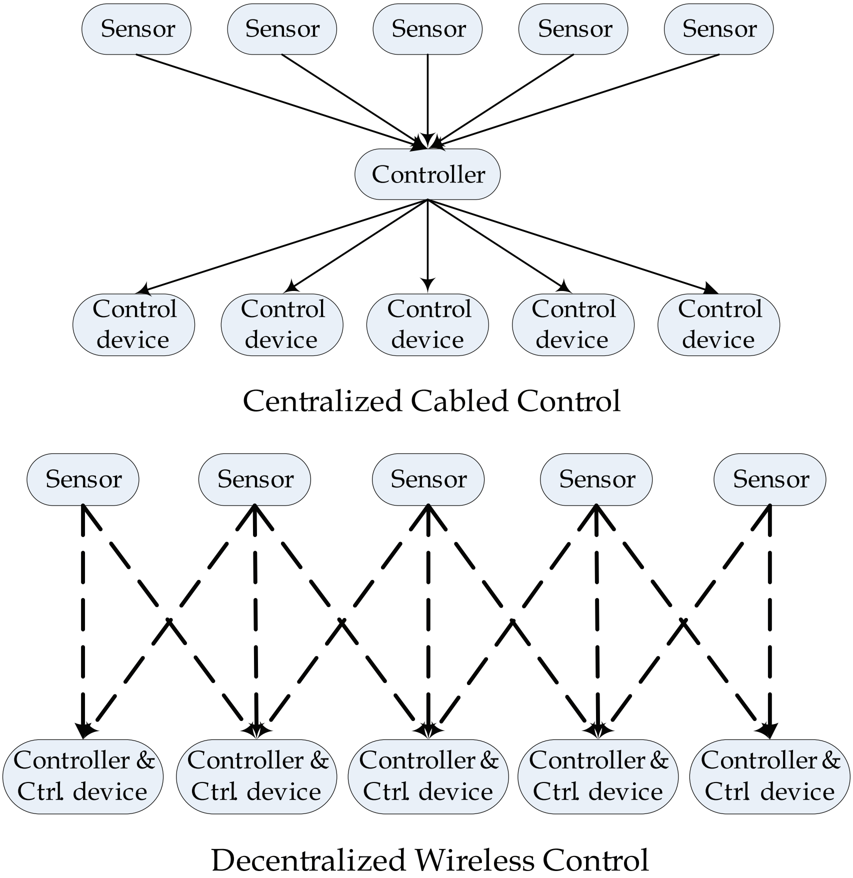
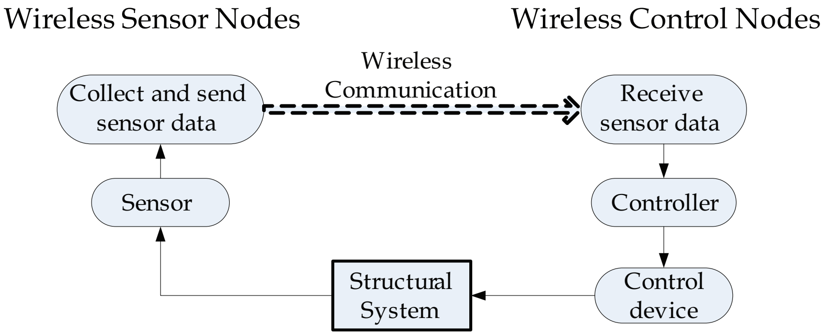


Fig. 10. Centralized and decentralized control systems

For application in wireless feedback structural control, real-time communication is important for system performance. Limited wireless communication range poses another challenge while instrumenting a large-scale structure with the wireless sensing and control system. Particularly, in discrete-time feedback control, a steady sampling time step and low communication latency are essential for the system performance. The feedback control loop designed for the prototype wireless sensing and control system is illustrated in Fig. 11(a), and the pseudo code implementing the feedback loop is presented in Fig. 11(b). As shown in the figures, sensing is designed to be clock-driven, while control is designed to be event-driven. The wireless sensing nodes collect sensor data at a preset sampling rate, and transmit the data during an assigned time slot. Upon receiving the required sensor data, the control nodes immediately compute control decisions and apply the corresponding command signals to the control devices. If due to occasional data packet loss, a control node doesn't receive the expected sensor data at one time step, the control node may use a projected data sample for control decisions, or doesn't take any action at this time step.

4.2 Communication protocol design for the wireless structural control system

Similar to the structural monitoring application, a reliable communication protocol must be properly designed for the wireless structural control system. Fig. 12 illustrates the communication state diagrams of a coordinator unit and other wireless units within a wireless sensing and control subnet. To initiate the system operation, the coordinator unit first broadcasts a start command '01StartCtrl' to all other sensing and control units. Once the start command and its acknowledgement '03AcknStartCtrl' are received, the system starts real-time feedback control operation, i.e. both the coordinator and other units are in State 2.



(a) Feedback control loop between the wireless sensing nodes and control nodes

```
        Wireless Sensing Nodes                      Wireless Control Nodes
            (Clock-driven)                              (Event-driven)
ITERATE {                                   ITERATE {
                                            IF (sensor data arrived on time)
Wait for the assigned time slot.            Compute control decisions.
                                            Apply control command signal.
Sample sensor data.                         ELSE
                                            Use projected data sample or no action.
Wirelessly transmit sensor data. ──►        Wait for the wireless sensor data.
}                                           }
```

(b) Pseudo code for the feedback control loop

Fig. 11. Illustration of the feedback control loop in a wireless decentralized control system

At every sampling time step, the coordinator unit broadcasts a beacon signal '02BeaconData' together with its own sensor data, announcing the start of a new time step. Upon receiving the beacon signal, other sensing units broadcast their sensor data following a preset transmission sequence, so that transmission collision is avoided. The wireless control units responsible for commanding the control devices receive the sensor data, calculate desired control forces, and apply control commands at each time step. In order to guarantee a constant sampling time step and to minimize feedback latency, timeout checking or retransmission is not recommended during the feedback control operation. This design is suitable for both centralized control and decentralized control.

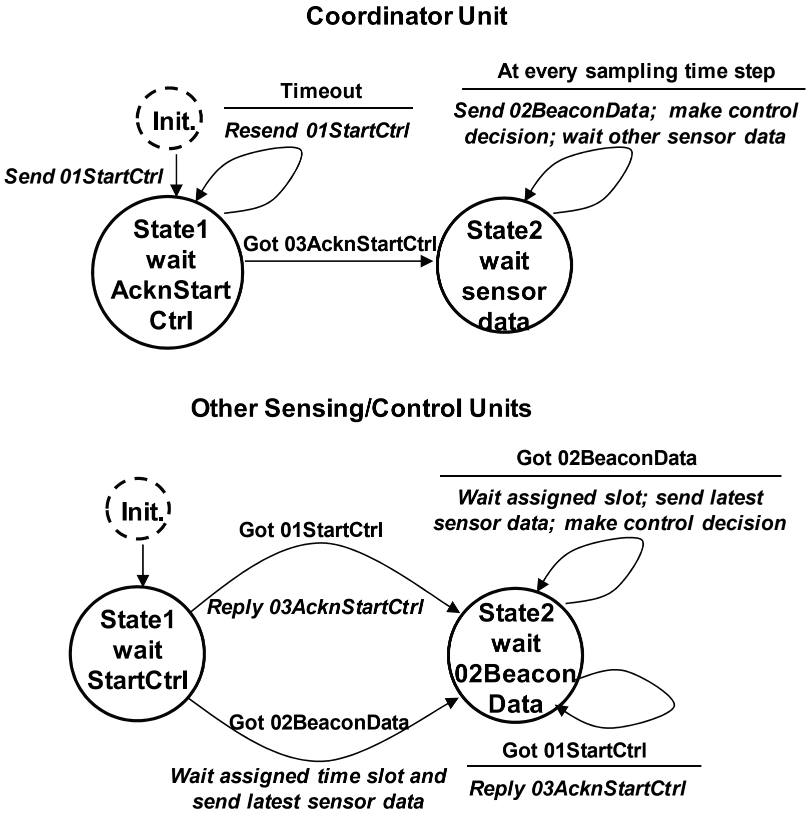
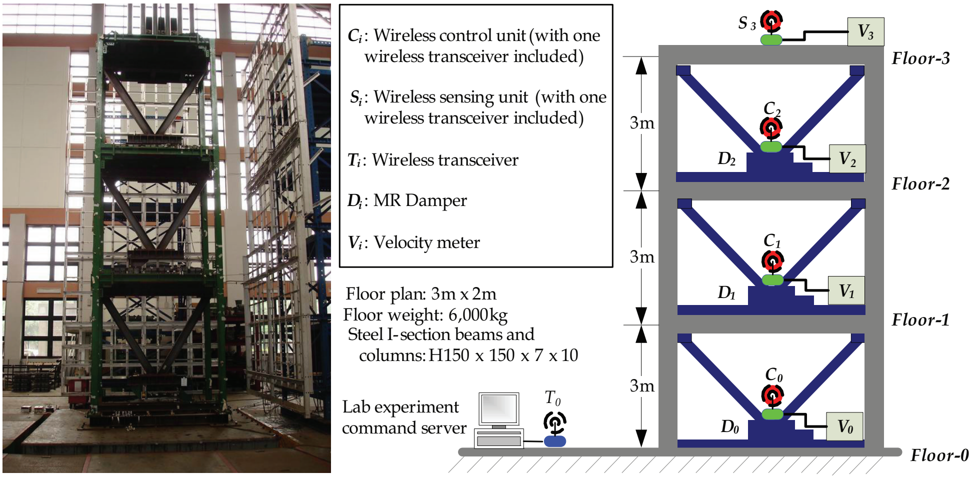

Coordinator Unit
Init.
Send 01StartCtrl
Timeout
Resend 01StartCtrl
State1 wait AcknStart Ctrl
Got 03AcknStartCtrl
At every sampling time step
Send 02BeaconData; make control decision; wait other sensor data
State2 wait sensor data
Other Sensing/Control Units
Init.
State1 wait StartCtrl
Got 01StartCtrl
Reply 03AcknStartCtrl
Got 02BeaconData
Wait assigned slot; send latest sensor data; make control decision
State2 wait 02Beacon Data
Got 02BeaconData
Wait assigned time slot and send latest sensor data
Got 01StartCtrl
Reply 03AcknStartCtrl

Fig. 12. Communication state diagram of a coordinator unit and other sensing/control units in one wireless subnet

For illustration purpose, a 3-story structure instrumented with the prototype wireless control system is shown in Fig. 13. The steel frame structure is designed and constructed by researchers affiliated with the National Center for Research on Earthquake Engineering (NCREE) in Taipei, Taiwan. The prototype wireless system consists of wireless sensors and controllers that are mounted on the structure for measuring structural response data and commanding MR dampers in real-time. Besides the wireless sensing and control units that are necessary for data collection and the operation of the control devices, a remote command server with a wireless transceiver is also included for experimental purpose. In a laboratory setup, the server is designed to initiate the operation of the control system and to log the data flow in the wireless network. To initiate the operation, the command server first broadcasts a start signal to all the wireless sensing and control units. Once the start command is received, the wireless units that are responsible for collecting sensor data start acquiring and broadcasting data at a preset time interval. Accordingly, the wireless units responsible for commanding the MR dampers receive the sensor data, calculate desired control forces, and apply control commands within the specified time interval.


C_i: Wireless control unit (with one wireless transceiver included)
S_i: Wireless sensing unit (with one wireless transceiver included)
T_i: Wireless transceiver
D_i: MR Damper
V_i: Velocity meter
Floor plan: 3m x 2m
Floor weight: 6,000kg
Steel I-section beams and columns: H150 x 150 x 7 x 10
Lab experiment command server
T_0
S_3
V_3
Floor-3
C_2
D_2
V_2
Floor-2
C_1
D_1
V_1
Floor-1
C_0
D_0
V_0
Floor-0
3m
3m
3m

(a) A 3-story test structure mounted on the shake table

(b) Deployment of the wireless sensors, controllers, and control devices

Fig. 13. Laboratory setup of the wireless structural control system

To coordinate the wireless transmissions during the feedback control, a pre-specified communication sequence should be observed by all the wireless units. For example, if all three wireless control units need velocity data from all the floors to compute control decisions, a communication sequence illustrated in Fig. 14 can be adopted by the prototype system. The control sampling step, which is 80ms in this example, is mostly decided by the total time required for transmitting all four data packets. For the 24XStream wireless transceiver adopted in the system, wireless transmission of each velocity measurement takes about 18ms. During every control time step, the wireless unit C_0 first samples the velocity data V_0 at its own floor, and then sends out the data together with a beacon signal to other wireless units. Upon receiving the beacon signal, units C_1, C_2, and S_3 sequentially broadcast their sensor data. Last, a period of 8ms is designed as a safety cushion for each control sampling time step, allowing certain randomness in the wireless transmission time. The control units C_0, C_1, and C_2 compute control decisions and apply actuation signals during the intervals of wireless transmissions.

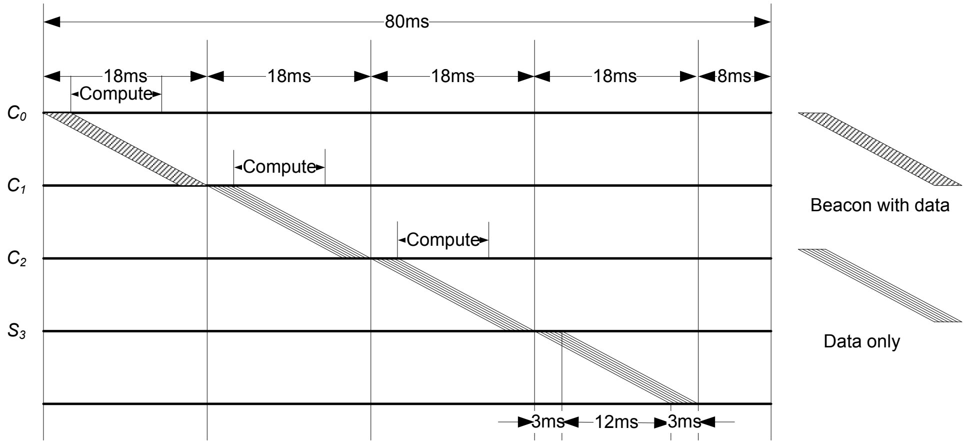


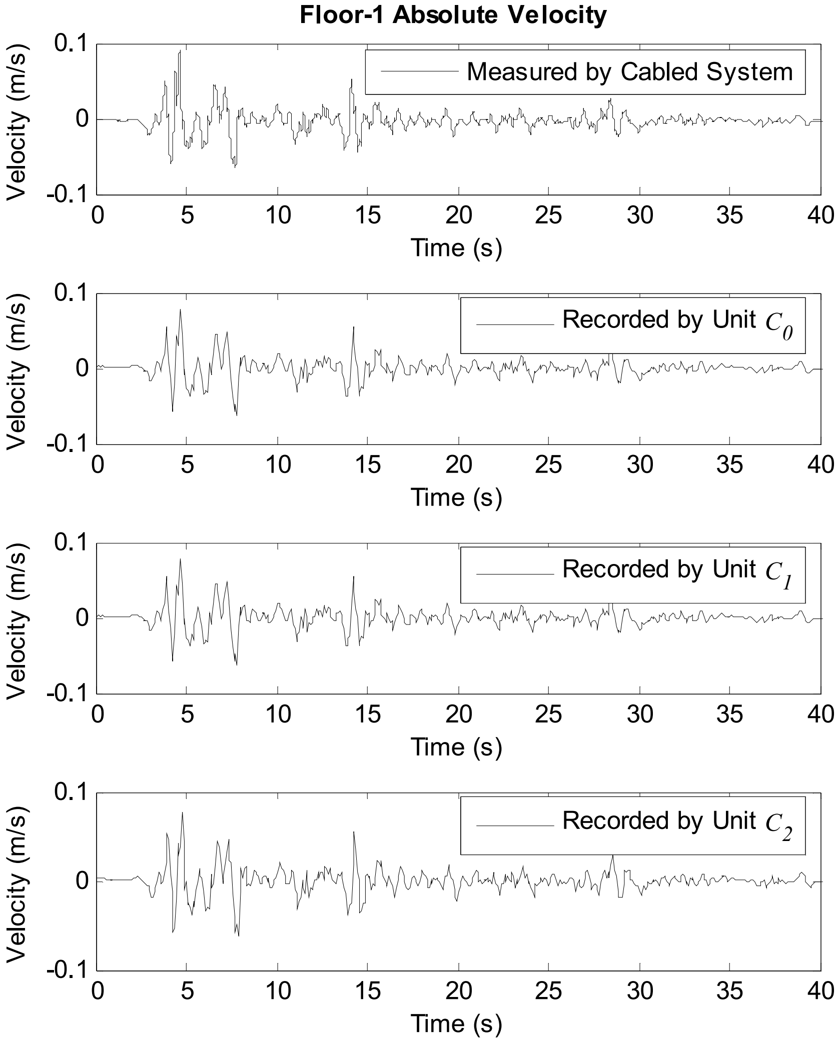
Fig. 14. Communication sequence in a wireless structural control network

4.3 Validation experiments for the wireless structural control system

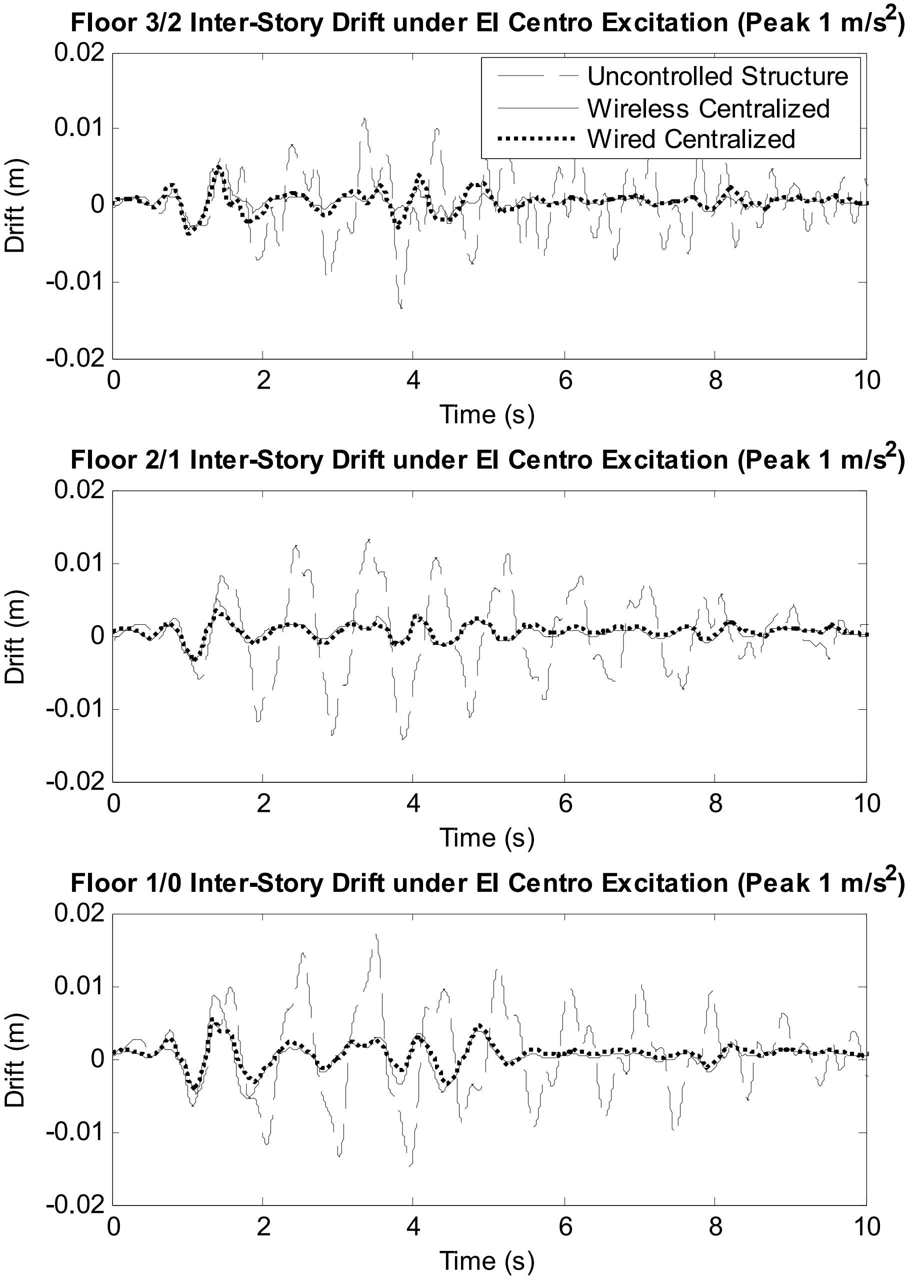
Validation experiments for the wireless control system were conducted at NCREE in Taipei, Taiwan, using the structure shown in Fig. 13. The floor plan of this structure is 3m × 2m, with each floor weight adjusted to 6,000 kg using concrete blocks; inter-story heights are 3m. The three-story structure is mounted on a 5m × 5m 6-DOF shake table. For this study, only longitudinal excitation in one degree of freedom is applied. Besides wireless sensors, a separate set of accelerometers, velocity meters, and linear variable displacement transducers (LVDT) are installed on each floor of the structure; this set of sensors are interfaced to a high-precision tethered data acquisition (DAQ) system native to the NCREE facility.

For this experimental study, three 20 kN MR dampers are deployed. Each damper is installed under a V-brace upon one of the three floors (Fig. 13b). The damping coefficients of the MR dampers can be changed by issuing a command voltage between 0V to 1.2V. This command voltage determines the electric current of the electromagnetic coil inside the MR damper, which, in turn, generates a magnetic field that sets the viscous damping properties of the MR damper fluid (Lin, *et al.* 2005). Two control systems, the wireless control system and a traditional wire-based control system, are installed in the test structure. For the wireless system, a total of four wireless sensors are installed to measure floor velocities (Fig. 13). Velocity feedback control algorithms presented in a previous paper are used by both the wired and the wireless control systems (Wang, *et al.* 2007b). In a centralized feedback pattern, real-time data from all sensors are required for making the control decisions for every MR damper. For this test structure, the wire-based system can achieve a sampling rate of 200Hz; as shown in Fig. 14, the wireless system can achieve a sampling rate of 12.5Hz.

In order to ensure that appropriate control decisions are computed by the wireless control units, one necessary condition is that the real-time velocity data used by the control units are reliable. Rarely experiencing data losses during the experiments, our prototype wireless sensor network proves to be robust. As reported by Lynch, *et al.* (2008), data losses less than 2% are experienced. Should data loss be encountered, the wireless control unit is currently designed to simply use the data sample from the previous time step. To illustrate the reliability of the velocity data collected and transmitted by the wireless units, Fig. 15(a) presents the Floor-1 time history data during a centralized wireless control test. The data is collected by both the wired DAQ system and the three wireless control units. During the test, unit C_1 measures the data from the associated velocity meter directly, stores the data in its own memory bank, and transfers the data wirelessly to units C_0 and C_2. After the test run is completed, data from all the three control units are sequentially streamed to the experiment command server, where the results are plotted as shown in Fig. 15(a). These plots illustrate strong agreement among data recorded by the three wireless control units and by the wired system using a separate set of velocity meters and data acquisition system. It is shown that the velocity data are not only reliably measured by unit C_0, but also properly transmitted to other wireless control units in real-time.



(a) Floor-1 absolute velocity data recorded by the cabled and wireless sensing systems



(b) Inter-story drifts of the structure with and without control

Fig. 15. Experimental time histories

The time histories of the inter-story drifts from the same centralized wireless control test are plotted in Fig. 15(b), together with the drifts of a centralized wired control test and a bare structure test when the structure is not instrumented with any control system (i.e. the MR dampers are not installed). The same ground excitation (1940 El Centro NS earthquake record scaled to a peak ground acceleration of 1m/s^2) is used for all the three cases shown in Fig. 15(b). The results show that both the wireless and wired control systems achieve considerable performance in mitigating inter-story drifts. Running at a much shorter sampling time step, the wired centralized control system achieves slightly better control performance than the wireless centralized system in terms of mitigating inter-story drifts.

To further study different decentralized schemes with different communication latencies, three wireless control architectures are compared: (#1) decentralized, (#2) partially decentralized, and (#3) centralized. Fig. 16 illustrates the information feedback pattern of each control architecture. The fully decentralized pattern (Wireless #1) specifies that when computing control decisions, the MR damper at each floor only needs the inter-story velocity difference at Story 1. The partially decentralized pattern (Wireless #2) specifies that the control decisions require inter-story velocity from a neighboring floor. Finally, the centralized pattern (Wireless #3) indicates all velocities relative to ground are required by the control decisions. Different information patterns result in different sampling frequencies for each control architecture. Compared with the centralized scheme, the advantage of a decentralized architecture is that fewer communication and data processing are needed at each sampling time step, thereby reducing sampling time step length. As shown in Fig. 16, the wireless system can achieve a sampling rate of 16.67Hz for partially decentralized control and 50Hz for fully decentralized control.

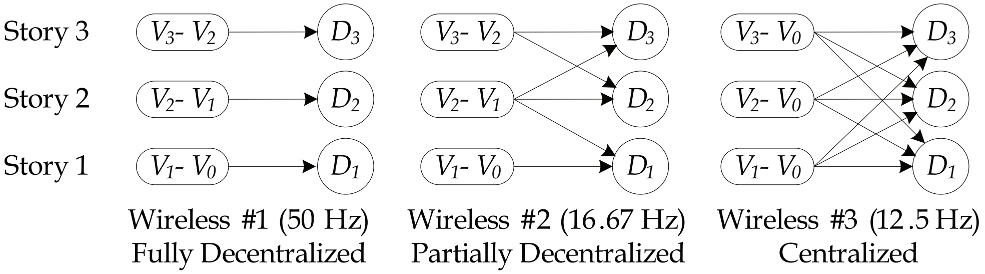


Fig. 16. Various decentralized and centralized information feedback

Fig. 17 shows the peak inter-story drifts and floor accelerations for the original uncontrolled structure and the structure controlled by the three different wireless schemes, as well as the wired centralized control scheme. The 1940 El Centro NS record is employed as the ground excitation, with peak ground acceleration scaled to 1m/s^2. Compared with the uncontrolled structure, all wireless and wired control schemes achieve significant reduction with respect to maximum inter-story drifts and absolute accelerations. Among the four control cases, the wired centralized control scheme shows good performance in mitigating both peak drifts and peak accelerations. This result is expected because the wired system has the advantages of lower communication latency and utilizes sensor data from all floors. The wireless schemes, although running at longer sampling steps, achieve control performance comparable to the wired system. For all three earthquake records, the fully decentralized wireless control scheme (Wireless #1) results in low peak inter-story drifts and the smallest peak floor accelerations at most of the floors. This result illustrates that in the decentralized wireless control cases, the higher sampling rate (achieved due to lower communication latency) potentially compensates for the lack of data from faraway floors.

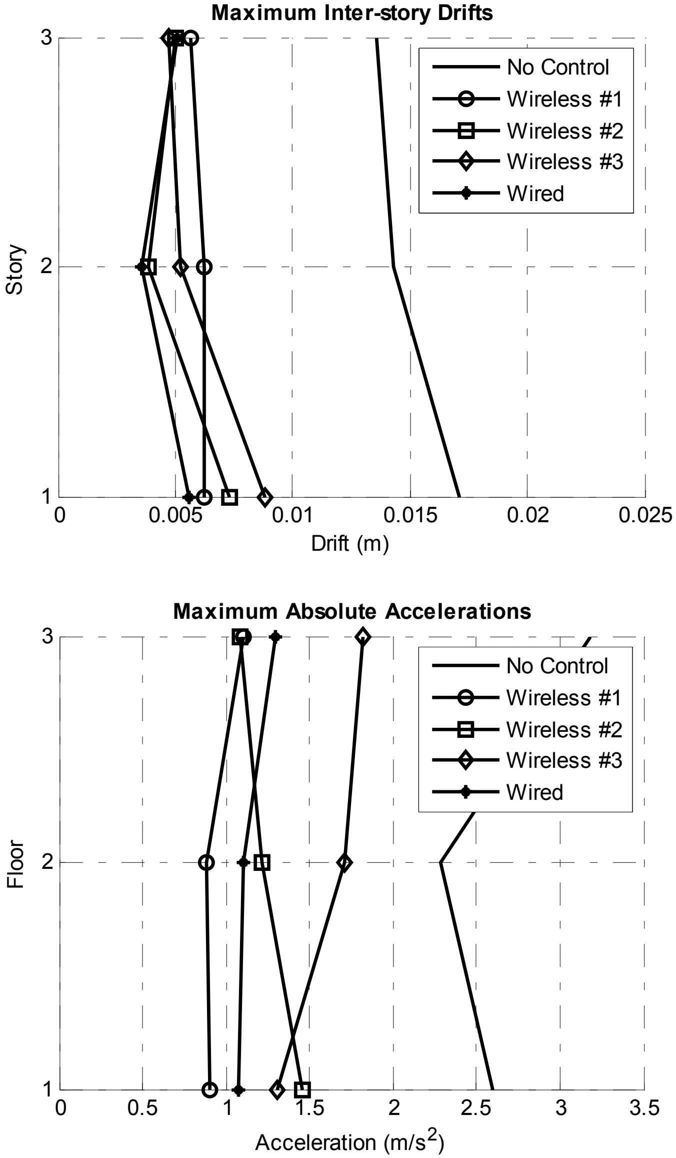


Fig. 17. Experimental results of different control schemes under 1940 El Centro NS earthquake excitation with peak ground accelerations (PGA) scaled to 1m/s^2

5. Summary and discussion

This chapter discusses the various issues of applying wireless sensor networks to modern smart structural technologies, including structural health monitoring and structural control. Autonomous wireless sensing and control units with embedded computing can serve as the building blocks of a smart structural system. For different structural applications, design concepts have been proposed to address the information constraints in a wireless sensor network, such as bandwidth, latency, range, and reliability. Robust communication protocol design for centralized and decentralized information architectures is proposed for efficiently managing the information flow in the wireless network. State machine concepts prove to be effective in designing simple yet efficient communication protocols for wireless structural sensing and control networks. Large-scale laboratory and field validation tests have been conducted to validate the efficacy and robustness of the information management schemes implemented in the wireless structural monitoring and control system. Most recently, the prototype wireless sensing system has been successfully tested for long-range measurement of low-amplitude and low-frequency vibrations at Canton Tower, a.k.a. Guangzhou TV and Sightseeing Tower, the world's tallest TV tower upon construction (Ni, *et al.* 2011).

A common trend in both structural monitoring and structural control application is the increasingly dense deployment of system nodes, i.e. sensors in a structural monitoring system, or sensors, controllers, and control devices in a structural control system. For example, in structural monitoring systems for cable-supported bridges, hundreds of sensors are often deployed for recording loading conditions and bridge responses (Wong 2004, Ko and Ni 2005, Çelebi 2006). Among many modern structural control systems, hundreds of semi-active hydraulic dampers have been installed in high-rise buildings (Kurino, *et al.* 2003, Spencer and Nagarajaiah 2003, Shimizu, *et al.* 2004). With rapid advancement in wireless sensor networks, there will be an inevitable trend of reduced system cost yet increased system nodal densities. Particularly in recent years, more and more large-scale wireless structural health monitoring (Lynch, *et al.* 2006, Kim, *et al.* 2007, Weng, *et al.* 2008, Whelan and Janoyan 2009, Rice, *et al.* 2010) and wireless structural control (Swartz and Lynch 2009, Wang and Law 2011) studies have been reported. Furthermore, researchers have started interesting exploration on mobile sensor networks, as the next-generation wireless sensor networks, for structural health monitoring applications (Zhu, *et al.* 2010). Such a mobile sensor network involves miniature autonomous mobile robots that carry wireless sensors and automatically move upon a large structure. In summary, it is believed that future monitoring and control systems will enjoy tremendous opportunities provided by the continuing advancements in wireless sensor technologies.

6. Acknowledgment

This research was partially funded by the National Science Foundation under grants CMS-9988909 and CMMI-0824977 (Stanford University), as well as CMMI-0928095 (Georgia Institute of Technology). The first author was supported by an Office of Technology Licensing Stanford Graduate Fellowship. Prof. Jerome P. Lynch at University of Michigan kindly provided insightful advices to the development of the prototype wireless sensing and control system. Prof. Chin-Hsiung Loh, Dr. Pei-Yang Lin, and Dr. Kung-Chun Lu at the National Taiwan University offered generous support for conducting the shake table experiments at NCREE, Taiwan. The authors would also like to express their gratitude to

Prof. Ahmed Elgamal and Dr. Michael Fraser of the University of California, San Diego, for their generous assistance throughout the field validation tests at Voigt Bridge.

7. References

Çelebi, M. (2006). Real-time seismic monitoring of the new Cape Girardeau Bridge and preliminary analyses of recorded data: an overview. *Earthquake Spectra*, Vol. 22, No. 3, pp. 609-630

Çelebi, M. (2002). *Seismic Instrumentation of Buildings (with Emphasis on Federal Buildings)*. Report No. 0-7460-68170, United States Geological Survey, Menlo Park, CA

Chang, F.-K. (Ed.) Structural Health Monitoring 2009: From System Integration to Autonomous Systems, *Proceedings of the 6th International Workshop on Structural Health Monitoring*, Lancaster, PA, USA, September 9-11, 2009

Cooklev, T. (2004). *Wireless Communication Standards : a Study of IEEE 802.11, 802.15, and 802.16*, Standards Information Network IEEE Press, New York

Doebling, S. W., Farrar, C. R. & Cornwell, P. J. (1997). DIAMOND: A graphical interface toolbox for comparative modal analysis and damage identification, *Proceedings of the 6th International Conference on Recent Advances in Structural Dynamics*, Southampton, UK, July 14 - 17, 1997

Farrar, C. R., Sohn, H., Hemez, F. M., Anderson, M. C., Bement, M. T., Cornwell, P. J., Doebling, S. W., Schultze, J. F., Lieven, N. & Robertson, A. N. (2003). *Damage Prognosis: Current Status and Future Needs*. Report No. LA-14051-MS, Los Alamos National Laboratory, Los Alamos, NM

Fraser, M., Elgamal, A. & Conte, J. P. (2006). *UCSD Powell Laboratory Smart Bridge Testbed*. Report No. SSRP 06/06, Department of Structural Engineering, University of California, San Diego, La Jolla, CA

Housner, G. W., Bergman, L. A., Caughey, T. K., Chassiakos, A. G., Claus, R. O., Masri, S. F., Skelton, R. E., Soong, T. T., Spencer, B. F., Jr. & Yao, J. T. P. (1997). Structural control: past, present, and future. *Journal of Engineering Mechanics*, Vol. 123, No. 9, pp. 897-971

Janssen, G. J. M. & Prasad, R. (1992). Propagation measurements in an indoor radio environment at 2.4 GHz, 4.75 GHz and 11.5 GHz, *Proceedings of IEEE 42nd Vehicular Technology Conference*, Denver, CO, May 10 - 13, 1992

Kim, S., Pakzad, S., Culler, D., Demmel, J., Fenves, G., Glaser, S. & Turon, M. (2007). Health monitoring of civil infrastructures using wireless sensor networks, *Proceedings of the 6th International Conference on Information Processing in Sensor Networks (IPSN '07)*, Cambridge, MA, April 25 - 27, 2007

Ko, J. M. & Ni, Y. Q. (2005). Technology developments in structural health monitoring of large-scale bridges. *Engineering Structures*, Vol. 27, No. 12, pp. 1715-1725

Kurino, H., Tagami, J., Shimizu, K. & Kobori, T. (2003). Switching oil damper with built-in controller for structural control. *Journal of Structural Engineering*, Vol. 129, No. 7, pp. 895-904

Lin, P.-Y., Roschke, P. N. & Loh, C.-H. (2005). System identification and real application of the smart magneto-rheological damper, *Proceedings of the 2005 International Symposium on Intelligent Control*, Limassol, Cyprus, June 27 - 29, 2005

Lynch, J. P. & Loh, K. J. (2006). A summary review of wireless sensors and sensor networks for structural health monitoring. *The Shock and Vibration Digest*, Vol. 38, No. 2, pp. 91-128

Lynch, J. P., Wang, Y., Loh, K. J., Yi, J.-H. & Yun, C.-B. (2006). Performance monitoring of the Geumdang Bridge using a dense network of high-resolution wireless sensors. *Smart Materials and Structures*, Vol. 15, No. 6, pp. 1561-1575

Lynch, J. P., Wang, Y., Swartz, R. A., Lu, K.-C. & Loh, C.-H. (2008). Implementation of a closed-loop structural control system using wireless sensor networks. *Structural Control and Health Monitoring*, Vol. 15, No. 4, pp. 518-539

MaxStream, Inc. (2004). *9XCite™ OEM RF Module Product Manual v1.1*. Lindon, UT

MaxStream, Inc. (2005). *XStream™ OEM RF Module Product Manual v4.2B*. Lindon, UT

Molisch, A. F. (2005). *Wireless Communications*, John Wiley & Sons, IEEE Press, Chichester, West Sussex, England

Ni, Y. Q., Li, B., Lam, K. H., Zhu, D., Wang, Y., Lynch, J. P. & Law, K. H. (2011). In-construction vibration monitoring of a super-tall structure using a long-range wireless sensing system. *Smart Structures and Systems*, Vol. 7, No. 2, pp. 83-102

Rappaport, T. S. & Sandhu, S. (1994). Radio-wave propagation for emerging wireless personal-communication systems. *Antennas and Propagation Magazine, IEEE*, Vol. 36, No. 5, pp. 14-24

Rice, J. A., Mechitov, K., Sim, S.-H., Nagayama, T., Jang, S., Kim, R., B. F. Spencer, J., Agha, G. & Fujino, Y. (2010). Flexible smart sensor framework for autonomous structural health monitoring. *Smart Structures and Systems*, Vol. 6, No. 5, pp. 423-438

Richardson, M. H. (1997). Is it a mode shape, or an operating deflection shape? *Sound and Vibration Magazine*, Vol. 31, No. pp. 54-61

Shimizu, K., Yamada, T., Tagami, J. & Kurino, H. (2004). Vibration tests of actual buildings with semi-active switching oil damper, *Proceedings of the 13th World Conference on Earthquake Engineering*, Vancouver, B.C., Canada, August 1 - 6, 2004

Sohn, H., Farrar, C. R., Hemez, F. M., Shunk, D. D., Stinemates, D. W. & Nadler, B. R. (2003). *A Review of Structural Health Monitoring Literature: 1996-2001*. Report No. LA-13976-MS, Los Alamos National Laboratory, Los Alamos, NM

Soong, T. T. (1990). *Active Structural Control: Theory and Practice*, Wiley, Harlow, Essex, England

Spencer, B. F., Jr. & Nagarajaiah, S. (2003). State of the art of structural control. *Journal of Structural Engineering*, Vol. 129, No. 7, pp. 845-856

Straser, E. G. & Kiremidjian, A. S. (1998). *A Modular, Wireless Damage Monitoring System for Structures*. Report No. 128, John A. Blume Earthquake Eng. Ctr., Stanford University, Stanford, CA

Swartz, R. A. & Lynch, J. P. (2009). Strategic network utilization in a wireless structural control system for seismically excited structures. *Journal of Structural Engineering*, Vol. 135, No. 5, pp. 597-608

Tweed, D. (1994). Designing real-time embedded software using state-machine concepts, *Circuit Cellar Ink*, (53), pp. 12-19.

Wang, Y., Lynch, J. P. & Law, K. H. (2005). Design of a low-power wireless structural monitoring system for collaborative computational algorithms, *Proceedings of SPIE, Health Monitoring and Smart Nondestructive Evaluation of Structural and Biological Systems IV*, San Diego, CA, March 9, 2005

Wang, Y. (2007). *Wireless Sensing and Decentralized Control for Civil Structures: Theory and Implementation*. PhD Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA

Wang, Y., Lynch, J. P. & Law, K. H. (2007a). A wireless structural health monitoring system with multithreaded sensing devices: design and validation. *Structure and Infrastructure Engineering*, Vol. 3, No. 2, pp. 103-120

Wang, Y., Swartz, R. A., Lynch, J. P., Law, K. H., Lu, K.-C. & Loh, C.-H. (2007b). Decentralized civil structural control using real-time wireless sensing and embedded computing. *Smart Structures and Systems*, Vol. 3, No. 3, pp. 321-340

Wang, Y. & Law, K. H. (2011). Structural control with multi-subnet wireless sensing feedback: experimental validation of time-delayed decentralized H_∞ control design. *Advances in Structural Engineering*, Vol. 14, No. 1, pp. 25-39

Weng, J.-H., Loh, C.-H., Lynch, J. P., Lu, K.-C., Lin, P.-Y. & Wang, Y. (2008). Output-only modal identification of a cable-stayed bridge using wireless monitoring systems. *Engineering Structures*, Vol. 30, No. 7, pp. 1820-1830

Whelan, M. J. & Janoyan, K. D. (2009). Design of a robust, high-rate wireless sensor network for static and dynamic structural monitoring. *Journal of Intelligent Material Systems and Structures*, Vol. 20, No. 7, pp. 849-863

Wong, K.-Y. (2004). Instrumentation and health monitoring of cable-supported bridges. *Structural Control and Health Monitoring*, Vol. 11, No. 2, pp. 91-124

Yao, J. T. P. (1972). Concept of structural control. *Journal of Structural Division, ASCE*, Vol. 98, No. 7, pp. 1567-1574

Zhu, D., Yi, X., Wang, Y., Lee, K.-M. & Guo, J. (2010). A mobile sensing system for structural health monitoring: design and validation. *Smart Materials and Structures*, Vol. 19, No. 5, pp. 055011

20

Extending Applications of Dielectric Elastomer Artificial Muscles to Wireless Communication Systems

Seiki Chiba and Mikio Waki
Chiba Science Institute
Wits Inc.
Japan

1. Introduction

Electro active polymers (EAPs) are used for actuators that can electrically control their motions to resemble those of actual muscles. Thus, they are called artificial muscles. In addition, since EAPs are often made of flexible materials, they have also come to be called "soft actuators" in recent years. There are many types of EAPs such as dielectric elastomers (Perline & Chiba, 1992a), ionic polymer-metal composites (Oguro et al., 1999), electroconductive Polymers (Otero & Sansinera, 1998), and ion polymer gels (Osada et al., 1992b). Figure 1 shows typical EAPs.

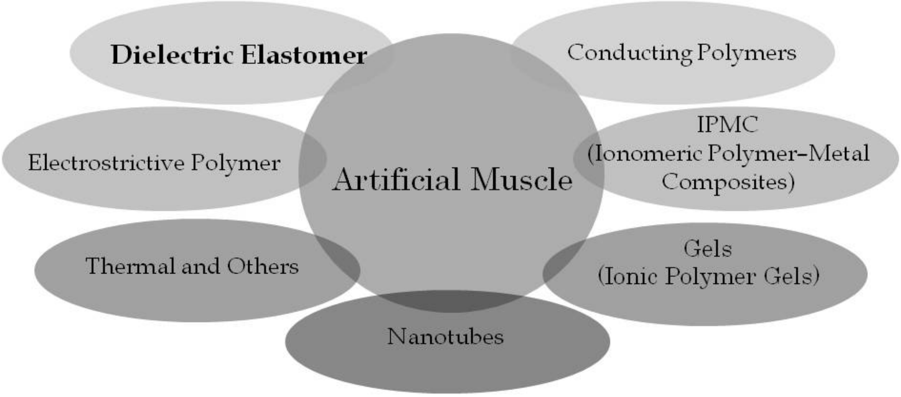


Fig. 1. Typical electro active polymers (EAPs)

EAP can be generally classified into two categories: elctrochemical polymers and fileld activated polymers (Kornbluh et al., 2004a). Electrochemical polymers use electrically driven mass transport of ions or electrically charged species to effect a charge in the shape (or vice versa). Field-activated polymers use an electric field to effect a shape change by acting directly on charges within the polymer (or vice versa). Each type of EAP has advantages and disadvantages for the application to wiress communications. Electrochemical polymers typically can exert relatively high pressures and can be driven by low voltages. However, they are relatively slow and limited in size (since they are dependent on molecular trnasport), require high current and relatively energy inefficient. They can operate best over a narrow range of temperatures and must often be kept moist (Kornbluh et al., 2004a). In contrast, field-activated polymers can be fast, efficient ,and relatively insensitive to temperature and humidity fluctuations. These polymers can operate at relatively high voltages and low currents, that usually requires additional voltage conversion components but makes the size and capacity of wires and interconnects lighter and less ctritical (Kornbluh et al., 2004a).

A type of field-activated EAP transducer that embodies the desirable proprties of polymer is dielectric elastomers (Pelrine et al., 2000).

Dielectric elastomer is a new transducer technology uses rubber like polymer (elastomer) as actuator materials. They have been gaining attention as technologies that have reached the practical use level as actuators and even as devices that can generate electricity efficiently (Chiba et al., 2008a).

The present paper examines the possibilities of frequency-variable antennas that utilize the actuator mode of dielectric-type artificial muscles, and sensor networks that utilize this electric generator mode (Chiba et al., 2007a; Chiba et al., 2008a).

2. Background on dielectric elastomer artificial muscles

Dielectric elastomer is a new smart material with characteristics and properties not seen in other materials. The basic element of dielectric erastomer is a very simple structure comprised of thin polymer films (elastomers) sandwiched by two electrodes made of a flexible and elastic material, and can operate as an electric control actuator.

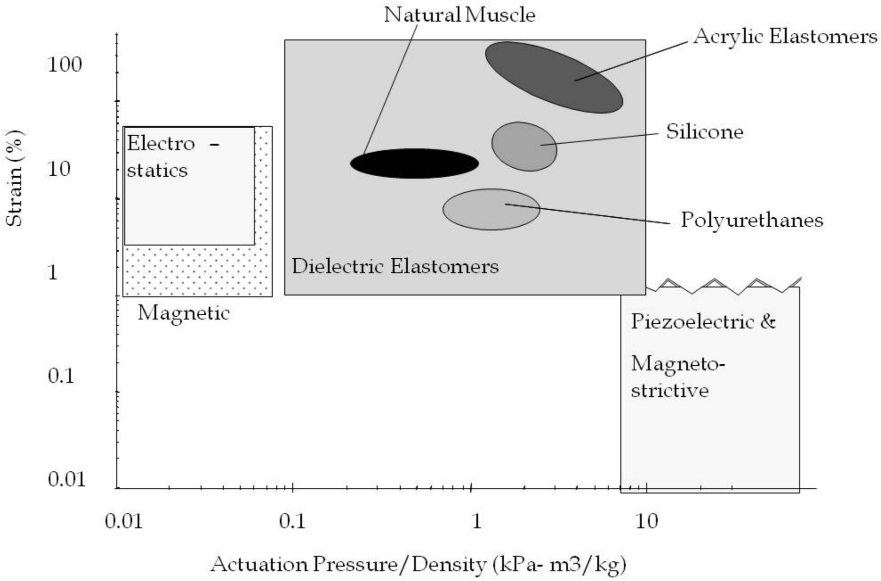


Fig. 2. Performance of dielectric elastomer is similar to that of natural muscle

Using a dielectric elrastomer actuator makes it possible to achieve a highly efficient transduction from electric energy into mechanical energy (the theoretical transduction efficiency is 80-90%, which translates into a considerable energy saving compared to other actuator technologies such as electric motors with gearboxes. At the material level, this material has fast speed of response (over 50,000 Hz has been demonstrated for small strains), with a high strain rate (up to 380% as shown in Photo 1), high pressure (up to 8 MPa), and power density of 1 W/g (for comparison, human muscle is 0.2 W/g and an electric motor with gearbox is 0.05 W/g) (Stanford et al., 2004a).

The energy density of dielectric elastomer has reached 3.4J/g, about 21 times that of single-crystal piezoelectrics and more than two oreders of magnitude greater than that of most commercial actuators (Pelrine et. al., 2000a). As can be seen in Figure 2, dielectric elastomers not only outperform existing actuator technologies in various areas but also are similar to natural muscle in that they fill the "actuator gap" between other actuation technologies, (Chiba, 2002). That is, dielectric elastomers have an actuation pressure/density that is bigger than that of electrostatic actuators and magnetic actuators, and cause strains that are bigger than that of piezo electric actuators and magneto strictive actuators.

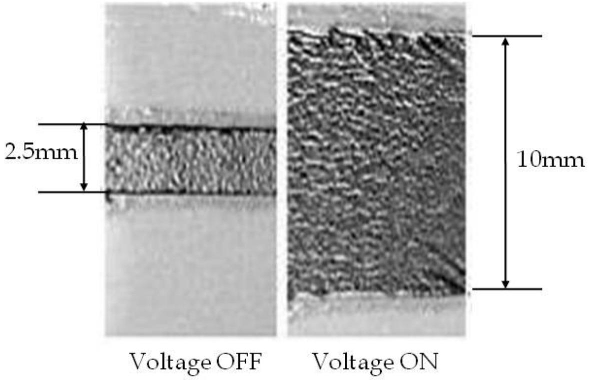


Photo 1. Acrylic elastomers showing 380% linear strain

2.1 Principle of operation of dielectric elastomers

Dielectric elastomer tranducers are based on the electromechanical response of an elastomeric dielectric film with compliant electrodes on each surface. Actuators based on dielectric elastomers technology operate on the simple principle shown in Figure 3. When a voltage is applied across the compliant electrodes, the polymer shrinks in thickness and expands in area.

The net volume change of the polymer materials that we investigate is small because of their high bulk moduli. Therefore, the electrodes must be compliant, to allow the film to strain. The observed response of the film is caused primarily by the interaction between the electrostatic charge on the electrodes. Simply put, the opposite charges on the two electrodes attract each other, while the like charges on the electrodes repel each other. Using this simple electrostatic model, we can derive the effective pressure produced by the electrodes on the film as function of the applied voltage. The pressure, ρ, is

$$\rho = \varepsilon \varepsilon_o E^2 = \varepsilon \varepsilon_o (V/t)^2 \tag{1}$$

where ε_o and ε are the permittivity of free space (8.85 x 10^{-12} F/m) and the relative permittivity (dielectric contact) of polymer, respectively; E is the applied electric field in V/m; V is the applied voltage; and *t* is the film thickness. The response of the polymer is functionally similar to that of electrostrictive polymers, in that the response is directly related to the square of the applied electric field,

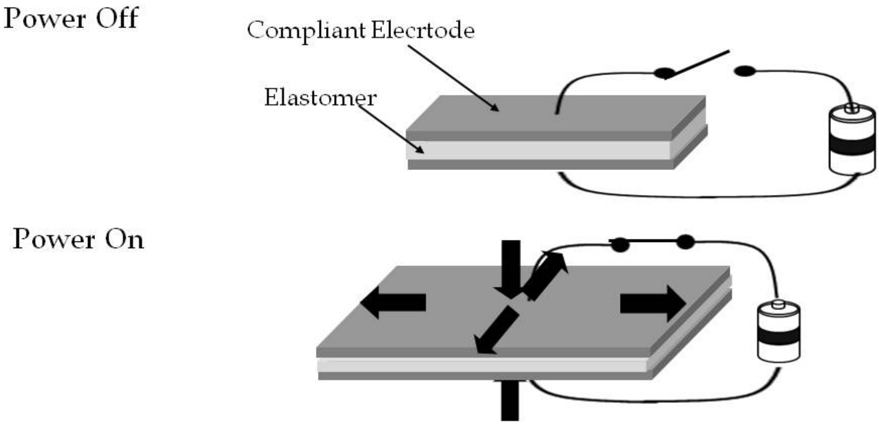


Fig. 3. Principle of operation of dielectric elastomers

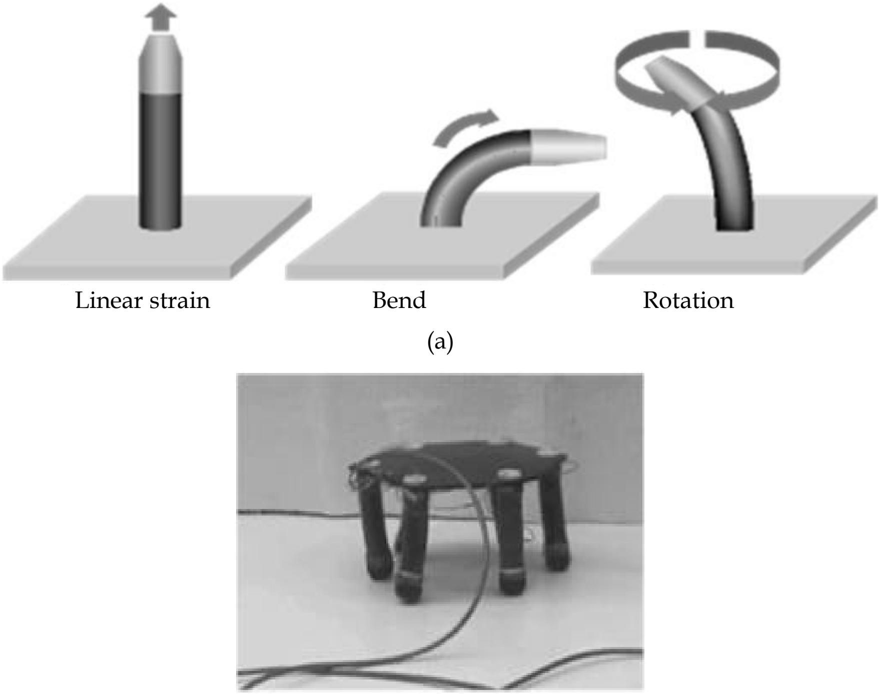
Two observation made from Equation 1 clarify the difference between Maxwell stress actuation and the use of conventional air-gap electrostatic actuators. First, ε for polymers is typically in the range 2-12, whereas for air ε is 1. Thus the actuation pressure is increased substantially via polymers rather than air the same electric field. Another difference is that typical air-gap actuators have an additional factor of 0.5 in their equivalent pressure expression, i.e., the polymers double the actuation pressure independent of the dielectric constant. The reason for this difference is that the polymers can stretch in area rather just contract in thickness. Polymers have two modes of converting electrical to mechanical energy. In contrast to polymers, air-gap actuators are typically made of rigid materials that can convert electrical to mechanical energy via only one mode of motion, such as the convergence of opposite electrodes.

Dielectric elastomers also have other advantages over air-gap electrostatic actuators, even though both are based on electrostatic force. Several polymers have been identified with breakdown strength of 300 MV/m or more in thin films, but breakdown strength this high are difficult to achieve consistently in air-gap electrostatic devices.

As mentioned above, the three effects, i.e., "two-mode coupling," "high dielectric susceptibility," and "high electric strength," greatly contribute to the actuation pressure of the dielectric elastomers.

2.2 Development summary of dielectric elastomer actuators

The elastomer has excellent workability which enables the shape design of devices with sizes from micrometers to several meters. Also, as elastomers are light and deform like rubbers, they can show flexible movements like bionic actions. They can express "flexible and natural feeling" which systems with motors cannot imitate. A wide array of proof-of-principle devices for use in leg robots (see Fig. 4), swimming robots, snakelike robots, compact inspection robots, geckolike robots for climbing up perpendicular walls or across ceilings, and flying robots, as well as in achieving compatibility with living organisms are currently developed (Stanford et. al., 2004b). The main feature of the dielectric elastomers is that they do not use any gears and cams, thus enabling high efficiency and safe and smooth driving even if the speed or direction of movement are suddenly changed.



(a)

(b)

Fig. 4. Biologically inspired robots powered by dielectric elastomer rolls (Pei et al, 2003; Chiba et al, 2006a). (a) Role Actuator Having 3-DOF (b) Application example to a robot: it enables sideways stepping like a crab without turning around, when it collides with wall

The 3-DOF actuator may be used as actuator for variable antenna of wireless communication device (see section 3 “Proof-of-principle experiment on a frequency-variable antenna utilizing the actuator mode of dielectric-type artificial muscles”).

Moreover, as this actuator has a wide dynamic range (DC to several tens of kHz), its applications to speakers and vibrational devices have been advanced (see Fig. 5) (Chiba et al., 2007a).

This device may be suitable for vibrators and speakers of wireless communication devices.

In addition, as there is a direct proportionality between the change in the capacitance and elongation of dielectric elastomer actuators, they can be used for pressure- and position-sensors (see Fig. 6). It may be possible to use the sensor function of dielectric elastomers to pick up electric waves for wireless communication devices.

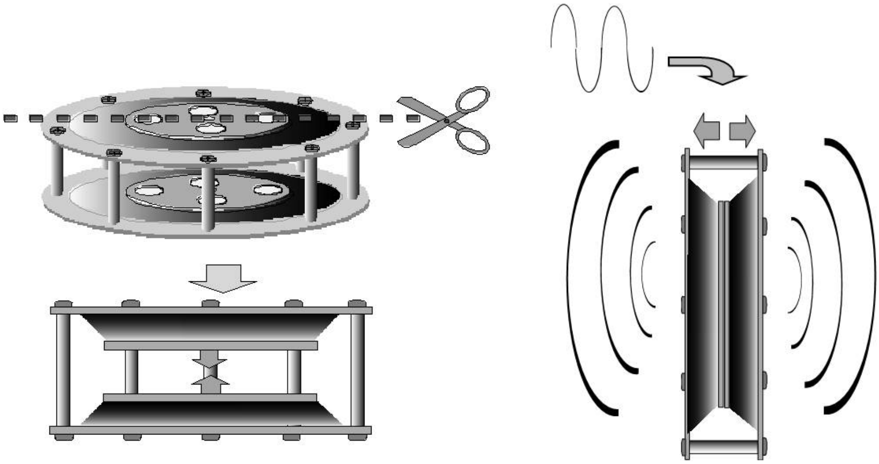


Fig. 5. Structure of speaker using dielectric elastomer (The black shaped part is dielectric elastromer) (Chiba et al., 2007a)

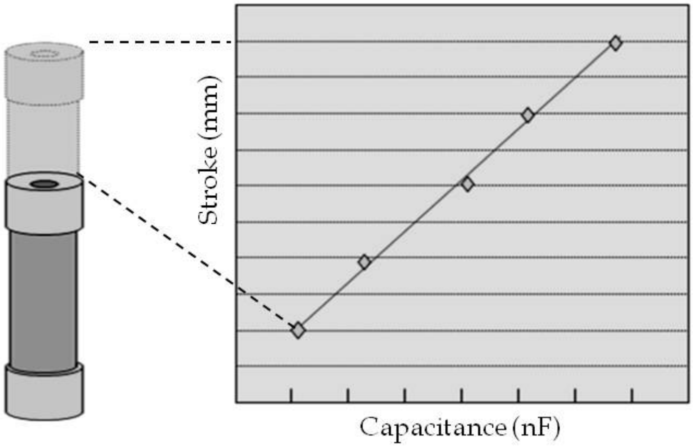


Fig. 6. Linear relation between capacitance and stroke of actuator (Kornbluh et al., 2004b)

3. Proof-of-principle experiment on a frequency-variable antenna utilizing the actuator mode of dielectric-type artificial muscles

The popularization of mobile telephones has brought wireless technology even closer to our daily lives. In recent years, improvements in integrated technology of electronic circuits and the increasing multi-functionality of mobile terminals have led to the inclusion of a multitude of diverse formats such as 3GPP, wireless LAN, Bluetooth, digital TV, etc., in single mobile communication devices. Since these communication formats all use different frequencies, it is necessary either to install a separate antenna for each wavelength, or use one antenna that can accommodate multiple frequencies.

Methods to create an antenna that is compatible for multiple frequencies include integrating antenna elements that can respond to multiple frequencies, and using an antenna that is shaped so that it can tune to a broad range of frequencies. The easiest method is to change the length of the antenna element, but because this changes the length of the antenna device, it requires equipment such as motors and gears. This makes it difficult to use as a small, lightweight frequency-variable antenna.

One way to resolve these problems may be to create a lightweight frequency-variable antenna with a simple structure by utilizing dielectric-type artificial muscles in the actuator part of a variable antenna.It may be possible to change the position of the reflection element and/or changing the length of dipolar- or monopolar antenna elements. Furthermore, by forming this structure onto polymers, it is possible to create a changeable-type planar antenna that can be installed in small, lightweight portable devices.

The present experiment corroborated the possibility of creating such variable-type antennas by using artificial muscle to change the length and tuning frequency of a monopolar antenna.

The variable-type monopolar antenna used in this experiment had a very simple structure. It was composed of a radial section that was attached to the dielectric artificial muscle actuator, and an antenna element section that was installed vertically on the core. (see Photo 2)

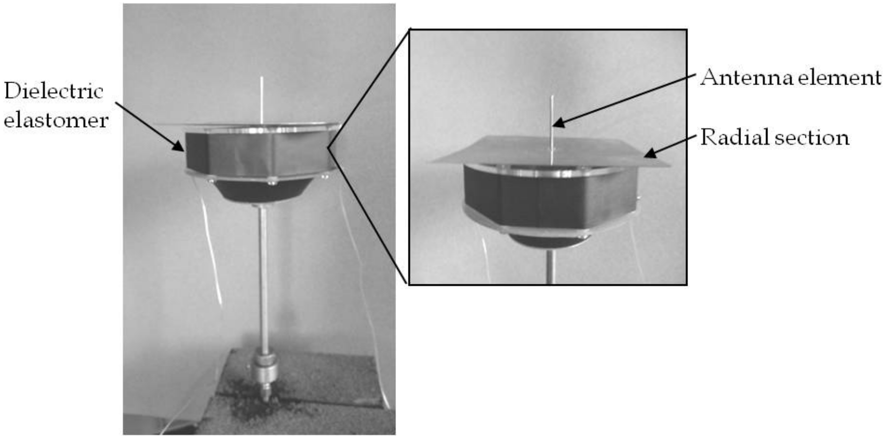


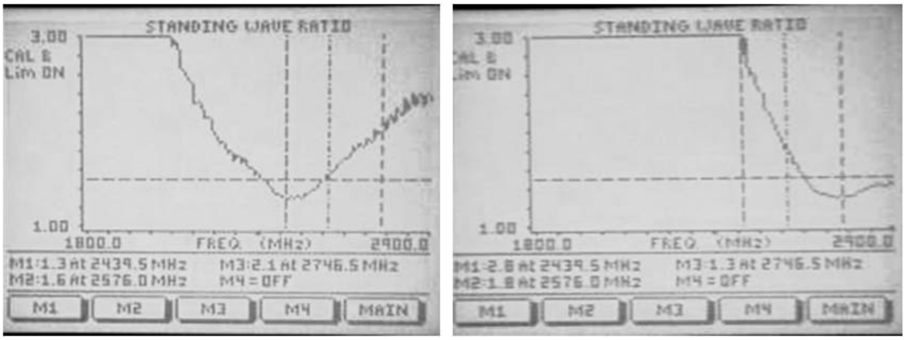
Photo 2. A frequency-variable antenna utilizing the actuator mode of dielectric elastomer artificial muscles

By changing the control voltage that was applied to the artificial muscle, a structure was created in which it was possible to change both the length of the antenna element part that was thrust out from the radial section and the tuning frequency.

In actuators that use dielectric artificial muscles, a thin (0.05 mm) elastomer film was attached to a 10 cm-diameter circular frame. By attaching two of these elastomers onto this frame, it became a diaphragm type with the cores of the elastomers attached to one another. The total weight, including the structural parts, was about 20 g.

The frequencies used in the experiment were in the 2.45 GHz band that is currently used in 3GPP, wireless LAN, and so on. The length L of the monopolar antenna element at a frequency of 2.45 GHz was 1/4 of the wavelength λ (122.4 mm), or 122.4/4 = 30.6 mm, and the changeable width of the actuator was 4 mm. This made it possible to change the tuning frequency within a range of about 300 MHz.

The change in the tuning frequency was confirmed by measuring V. S. W. R. (Voltage Standing Wave Ratio) using a network analyzer (Photo 3).



(a) Before change (b) At the time of the maximum change

Photo 3. Measurement of V. S. W. R. (The setting frequency range of a network analyzer: start frequency, 1.8 GHz and stop frequency, 2.9 GHz)

In this experiment, a diaphragm actuator for artificial muscle speakers was used, but this system was not smart, because the muscle part was too large. However, since the purpose of this experiment was to make the resonant frequency of a non-directional antenna variable by changing the length of the antenna element, a monopole antenna, which has the simplest structure, and artificial diaphragm muscles were used.

In our next experiment, we plan to change the direction of electric wave radiation by varying the installation angle of a directional antenna with roll-type artificial muscles.

In another words, the plan call for conducting an experiment to vary the directivity inside the vertical face of the antenna by making a model (ground plane) antenna by changing the wire in the radial part, and enabling the angle of attachment to the radial part to be changed by the roll-type artificial muscle. If such a variable antenna can be put to practical use, then it might be possible to create a system where the antenna can automatically be varied to match a more optimal electric wave environment, and even a small amount of electric power can be used to construct a suitable electric wave environment.

Furthermore, plans are being drawn for conducting an experiment on a planar antenna whose directivity and tuning frequency can be changed by using the dielectric-type artificial muscle to transform the antenna formed on the polymer. In the near future, by using variable antennas whose shape changes to match the use in mobile telephones, personal computers, etc., it may be possible to create a pleasant wireless communications environment with just a little bit of electrical power.

4. Sensor network that utilizes the power generation mode of a dielectric elastomer artificial muscle

Another working mode of the dielectric elastomer artificial muscle is the power generation mode. This is operatively the opposite of the actuator function. By adding external power to the dielectric type artificial muscle, the shape can be changed, and the increased static electrical energy produced therefrom can generate electricity. Since this power generation phenomenon is not dependent on the speed of transformation, its power generation device can generate electric energy by utilizing natural energies such as up-and-down motions of waves, slowly flowing river water, human and animal movements, and vibration energies produced from vehicles and buildings.

4.1 Principal of the power generation mode

The operation principle in the generator mode is the transformation of mechanical energy into electric energy by deformation of the dielectric elastomer (Ashida et al,:2000b). Functionally, this mode resembles piezoelectricity, but its power generation mechanism is fundamentally different. With dielectric elastomer, electric power can be generated even by a slow change in the shape of dielectric elastomer, while for piezoelectric devices impulsive mechanical forces are needed to generate the electric power. Also, the amount of electric energy generated and conversion efficiency from mechanical to electrical energy can be greater than that from piezoelectricity (Chiba et al,. 2007a). Fig.7 shows the operating principal of dielectric elastomer power generation.

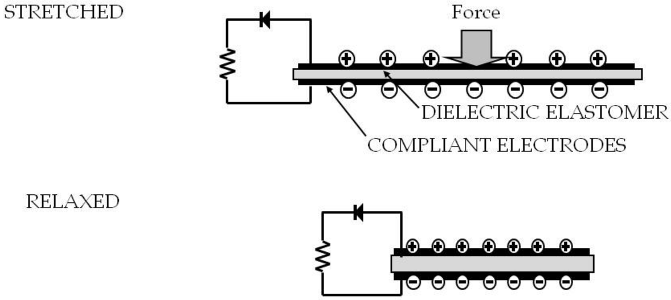


Fig. 7. Operating principle of dielectric elastomer power generation

Application of mechanical energy to dielectric elastomer to stretch it causes compression in thickness and expansion of the surface area. At this moment, electrostatic energy is produced and stored on the polymer as electric charge. When the mechanical energy decreases, the recovery force of the dielectric elastomer acts to restore the original thickness and to decrease the in-plane area. At this time, the electric charge is pushed out to the electrode direction. This change in electric charge increases the voltage difference, resulting in an increase of electrostatic energy.

$$C = \varepsilon_0 \varepsilon A/t = \varepsilon_0 \varepsilon b/t^2 \tag{1}$$

where ε_0 is the dielectric permittivity of free space, ε is the dielectric constant of the polymer film, A is the active polymer area, and t and b are the thickness and the volume of the polymer. The second equality in Equation (1) can be written because the volume of elastomer is essentially constant, i.e., $At = b =$ constant.

The energy output of a dielectric elastomer generator per cycle of stretching and contraction is

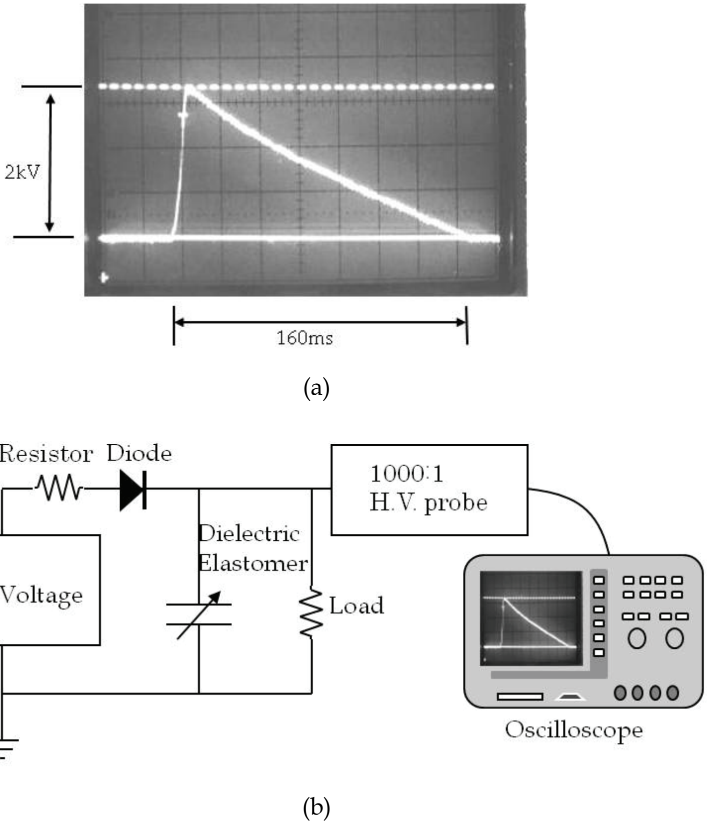
$$E = 0.5 C_1 V_b{}^2 (C_1/C_2 - 1) \tag{2}$$

where C_1 and C_2 are the total capacitances of the dielectric elastomer films in the stretched and contracted states, respectively, and V_b is the bias voltage.

Considering then changes with respect to voltages, the electric charge Q on a dielectric elastomer film can be considered to be constant over a short period of time and in the basic circuit. Since $V = Q/C$, the voltages in the stretched state and the contracted state can be expressed as V_1 and V_2, respectively, and the following equation is obtained:

$$V_2 = Q/C_2 = (C_1/C_2)\ (Q/C_1) = (C_1/C_2)\ V_1 \tag{3}$$

Since $C_2 < C_1$, the contracted voltage is higher than the stretched voltage, corresponding to the energy argument noted above. The higher voltage can be measured and compared with predictions based on the dielectric elastomer theory. In general, experimental data based on high impedance measurements are in excellent agreement with predictions. When the conductivity is assumed to be preserved in the range of electric charging, Q remains constant.



(a)

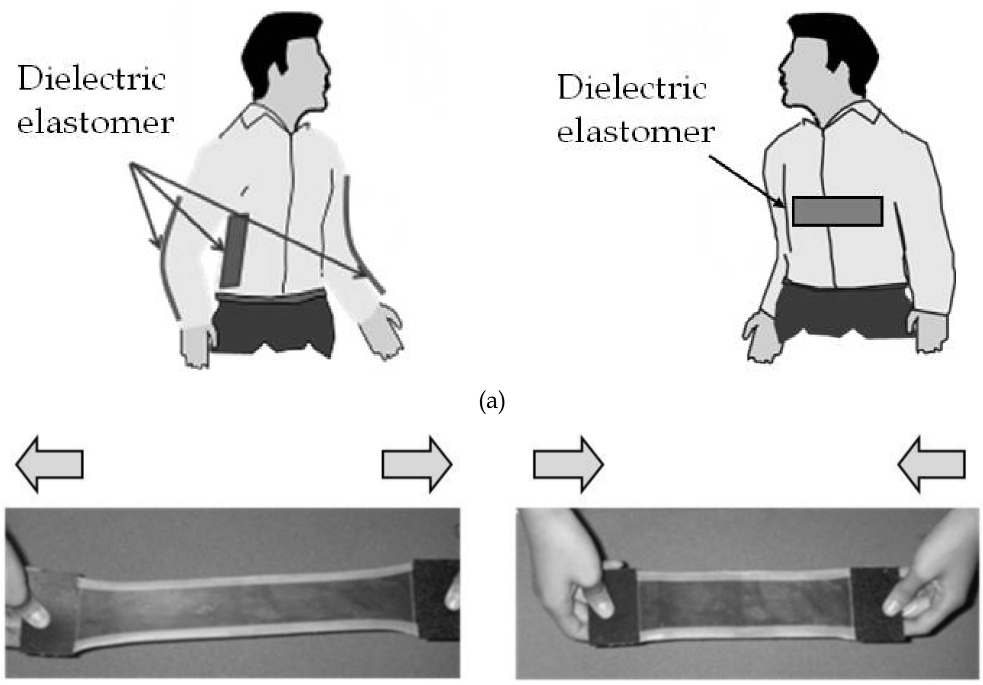
(b)

Fig. 8. Voltage for compression of dielectric elastomer and measurement circuit. (a) Typical scope trace from contraction of dielectric elastomer. Voltage spike occurs at contraction and gradually back to (stretched) voltage due to load resistance. (b) Measurement circuit of generated energy

Figure 8(a) shows a typical scope trace from contraction of dielectric elastomer. Figure 8(b) shows a simplified circuit for oscilloscope measurement of voltage. The voltage peak generated for one cycle is typically on the order of a few ms to several tens of ms for a piezoelectric element. However, in the case of dielectric elastomer, the peak width is on the order of 150-200 ms or longer (Chiba et al., 2008a). The long power-generation pulse duration of dielectric elastomer can allow for the direct use of generated energy for activities such as lighting LEDs. This can even power wireless equipment that is evolving today at a rapid pace. In continuous cyclical motions, it is easy to continuously obtain electrical energy by using flat and smooth circuits, even with gentle kinetic energy below a few Hz (Chiba et al., 2007b)

4.2 Application of dielectric elastomer generator to wireless communication system

In a power generation experiment, a thin artificial muscle film (25 cm long x 5cm wide, weight about 0.5 g) attached a human arm was able to generate 20 mJ of electrical energy with one arm movement. It is also possible to make them generate electricity putting up dielectric elastomers besides the arm to the side and the chest of the body (See Fig. 9a).



(a)

Streched state

Relaxed states

(b)

Fig. 9. Harvesting energy system from human body. (a) Conceptual rendering of dielectric elastomers put up to side and chest of arm and body: (b) Stretched state of dielectric elastomer (left) and Relaxed state of the elastomer (right)

Furthermore, in an experiment using different power generation equipment, artificial muscle film attached to the bottom of a shoe was verified to generate electricity when the artificial muscle was distorted while walking. When an adult male took one step per second, one shoe was able to produce about 1 W of electrical power. (Harsha et al., 2005)

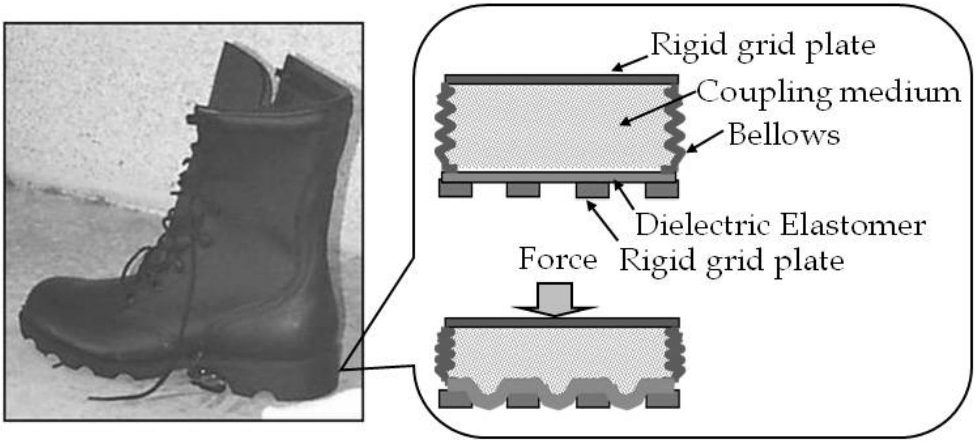


Fig. 10. Shoe generator

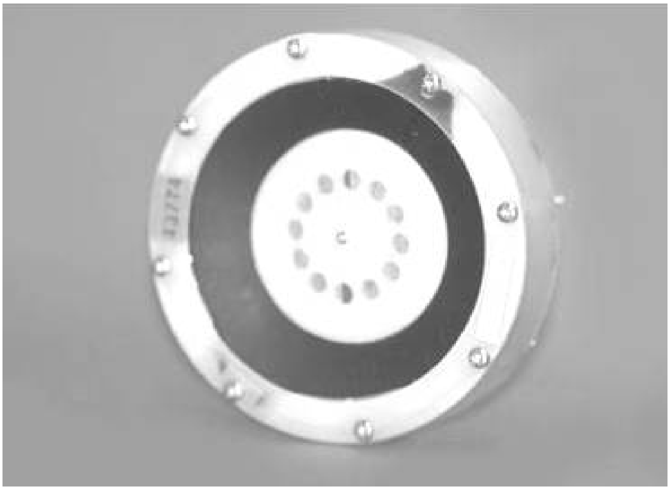
This confirmed that by utilizing human movement, enough electrical power could be obtained to recharge batteries for mobile telephones and similar devices (Chiba et al., 2008). In addition, electrical energy from the movements of animals could be used to construct livestock management systems. Other applications of animal-generated energy being investigated include scientific surveys of ecosystems of migratory birds and fish, among others.

In an experiment using a diaphragm actuator, electrical power output of about 0.12 – 0.15W was obtained by pressing the center of a roughly 1 g, 8 cm-diameter EPAM a few millimeters one time per second (Chiba et al., 2007a). Using the same equipment, the electric power generated was able to illuminate 6 LEDs, and by combining this with a wireless system, it became possible to turn a device on and off from a remote location.

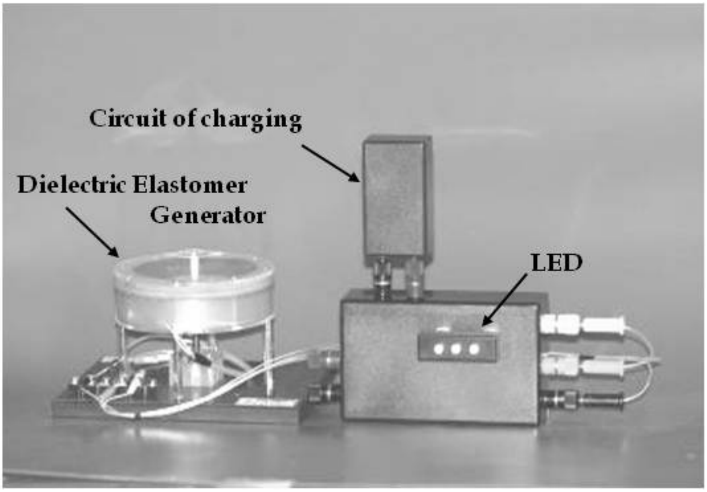
In such ways, dielectric elastomer artificial muscles can supply electrical power only when mechanical energy is obtained, and it is possible to simultaneously act as a switch that detects power sources and motion. Consequently, it may possible to easily create wireless networks, with simple components that do not require batteries (Chiba et al., 2007a).

In recent years, global warming and accompanying abnormal weather have begun to have an impact on our daily lives. To protect ourselves from the disasters brought about by abnormal weather, it is important to thoroughly understand the current situation, that is, how the global environment is changing.

The monitoring of the global environment has been done by various countries on their own, but to monitor environmental changes on a global scale it will be necessary to build wide-ranging sensor networks. One of the major issues with that, however, is that there is no good method for obtaining electrical energy for running this system. Presently, many if not most of these sensor systems are powered by solar batteries, but in some locations and during some seasons the daylight hours are extremely short, and in maritime and desert areas salt and dust can dramatically reduce the electrical output. All this makes it difficult to maintain a stable sensor system.



(a)



(b)

Photo 4. Small scale power generation device. a) Cartridge of used for small generator. The black ring-shaped part is dielectric elastomer. b) A power of approximately 0.12 W can be generated, by pushing the central part of dielectric elastomer by 3- 4 mm once a second

As one way of resolving these issues, power generation systems that utilize artificial muscles to generate power through transformation alone are attracting attention. Already, experiments using wave power to generate electricity have been able to produce a few watts of electrical energy with small artificial muscle power generation equipment loaded onto weather observation buoys, (see photo 6 and fig. 11) and this has also been confirmed to recharge batteries (Chiba et al., 2009).

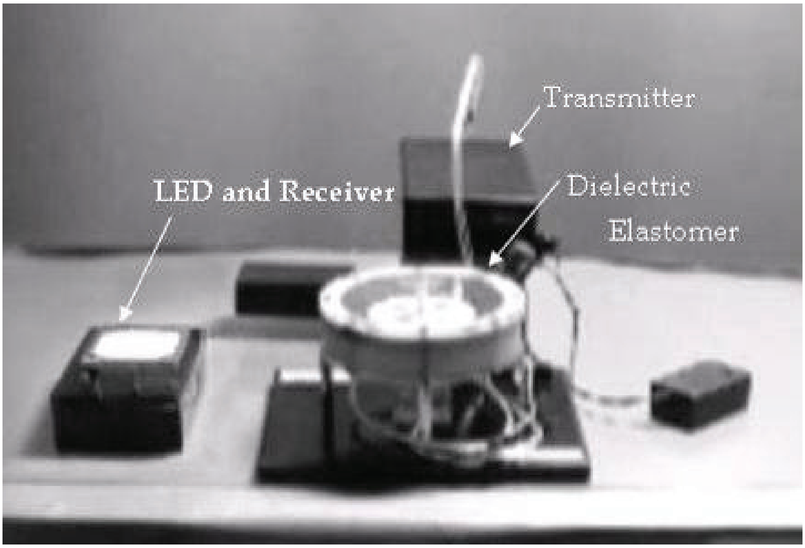


Photo 5. Small scale power generation device & LED controlled by wireless signals

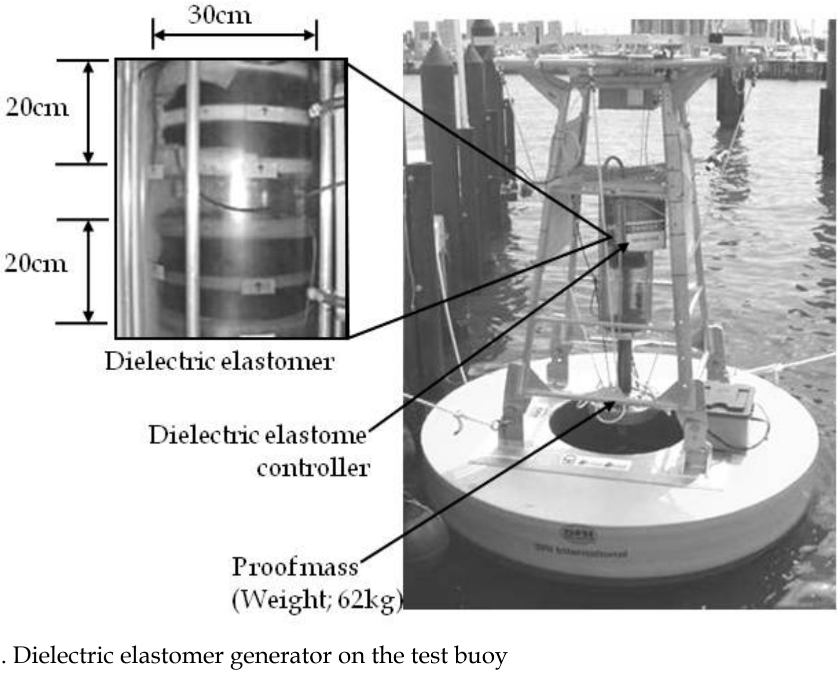


Photo 6. Dielectric elastomer generator on the test buoy

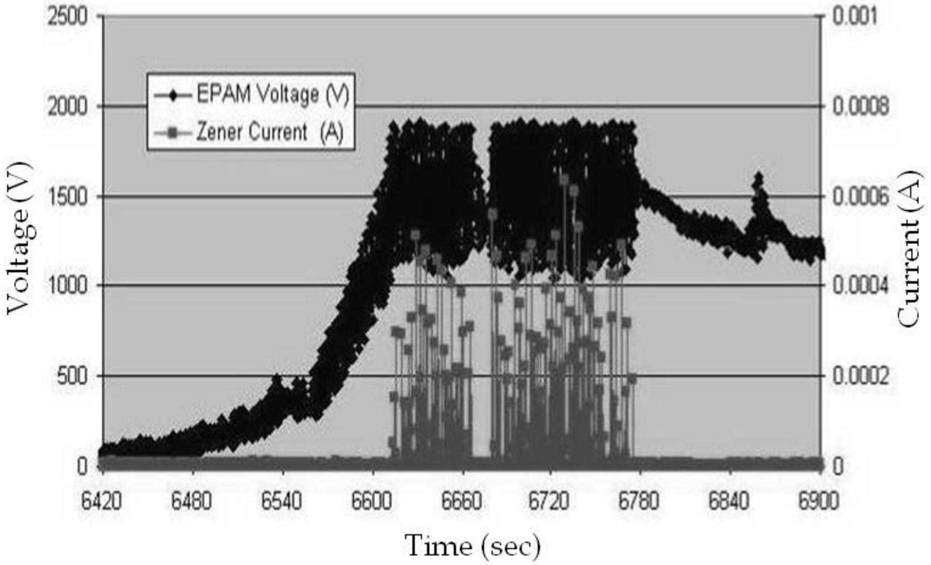


Fig. 11. Electricity generated by ten centimeter-high waves

In other experiments (see photo 7), electrical energy has been obtained from flowing water in a laboratory (Chiba et al., 2007a). The flow of water rotates the water-mill, and the rotational motion induces the deformation of the dielectric elastomer to generate electrical energy. Figure 12 shows the conceptual rendering of water mill generator using dielectric elastomer (Chiba et al., 2007a).

Furthermore, the results of simulations based on conceptual designs of flag-type power generation equipment using artificial muscles have indicated that there is little loss from the fluttering of flags and that it is possible to generate electric power with a high rate of efficiency (Chiba et al., 2007b).

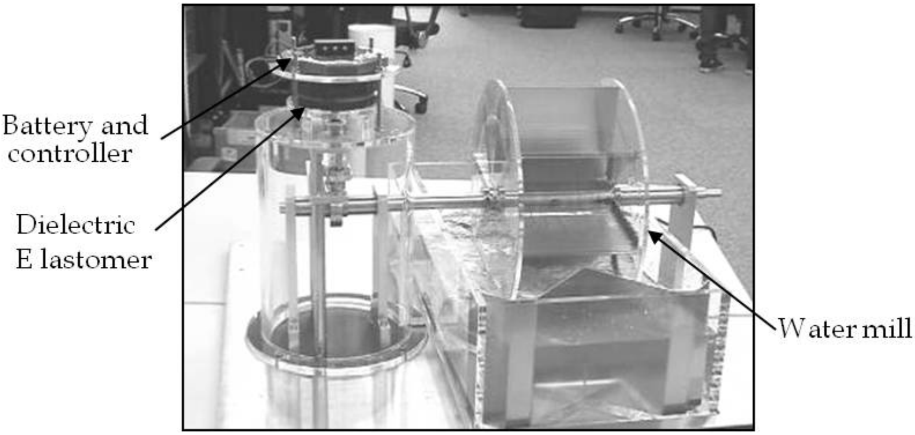


Photo 7. Water mill generator using dielectric elastomer

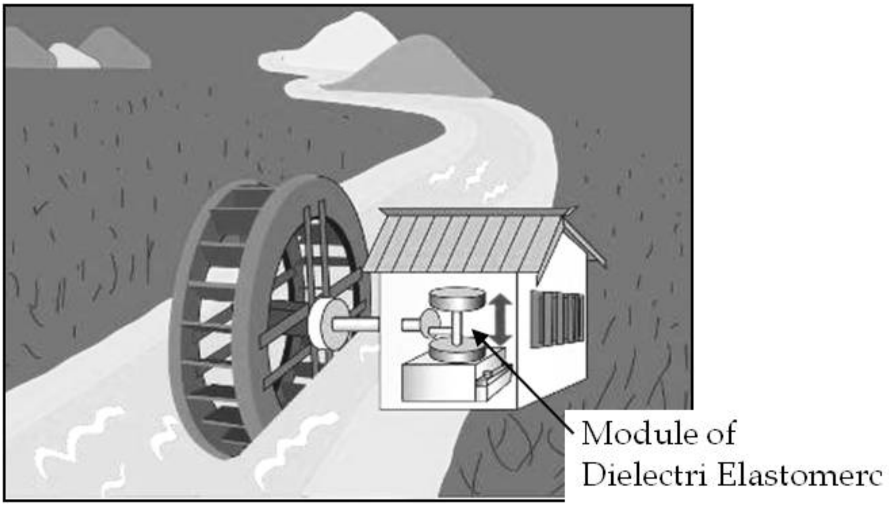


Fig. 12. Conceptual rendering of water mill generator using dielectric elastomer (Chiba et al., 2010)

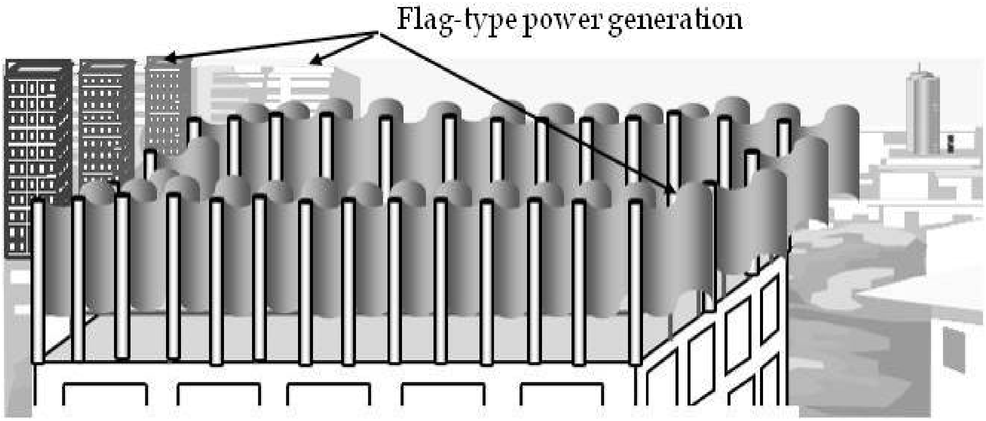


Fig. 13. Conceptual rendering of flag-type power generation

4.3 Analysis of power generation cost

Even without dielectric elastomer technology, ocean wave power is beginning to flourish in several countries. These ocean wave power systems typically use hydraulic pistons that are pumped by the wave action. The hydraulic fluid flows through a transmission and then a turbine to spin a rotary electromagnetic generator. When these systems are successfully developed for commercial use, the unit price of a power generation of kWh is estimated to be about 20 US Cents (Chiba et al., 2008b). These wave power systems are typically designed for ocean waves exceeding 2 - 3 m in height. At significantly smaller wave heights, the systems become less economically attractive (Miyazaki et al, 2007).

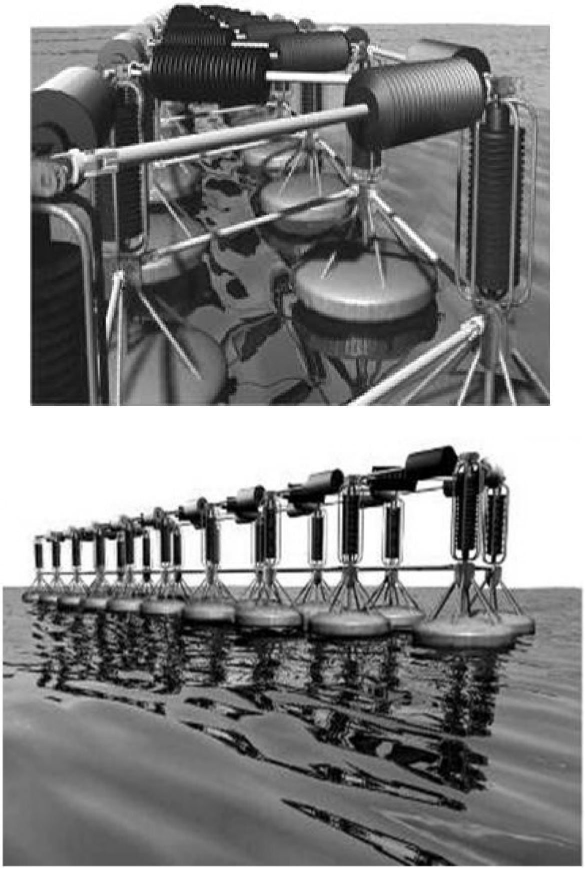


Fig. 14. Conceptual rendering of wave power generator system using dielectric elastomers

Because of its simplicity, efficiency, and size scalability, we believe that dielectric elastomer-based wave generator systems can be attractive not only for large wave applications but for many applications where the waves are much smaller. An estimate based on data from our sea trial demonstration experiments has shown that even in seas where the wave height is only 1 m throughout the year (e.g., the sea close to Japan), if there are spaces of approximately 500 m in length and 10 m in width, the establishment of a sea-based facility generating 6 MW of power is possible (Chiba et al., 2008b). This is a useful amount of power, be it for general use or for providing energy for nearby residential or industrial needs. The ability to produce the power where it is needed can eliminate the losses and costs associated with power transmission over long distances and make wave power even more attractive. The power generation efficiency estimated on the basis of the data obtained from in-tank experiments in 2006 (Chiba et al, 2006b) and ocean demonstration experiments in 2007 (Chiba et al, 2008a) and 2008 (Chiba et al, 2009) is approximately 19 US cents/kWh. In the near future, we expect that the electric power generation per unit mass or volume of dielectric elastomer material can double, and that the expected power generation cost per kilowatt-hour is 5 - 7.5 US cents. This value is comparable to that for fossil fuel thermal power plants. Of course, the wave power systems have the additional benefit of not releasing any pollution or greenhouse gasses.

5. Future of dielectric elastomer systems to wireless communication

The variable antenna technologies with artificial muscles have high expectations to apply to not only data communications for mobile phones and personal computers but also wireless sensor systems which monitor various data concerning weather conditions and environments. In the future, the combination of these artificial muscle power-generating systems with various sensing systems will make it possible to conduct sensing on a global scale, and may even make a significant contribution to the creation of systems that will protect human lives from natural disasters that have so far been difficult to predict.

Various power generating systems can be set up in each place on the earth as shown in Figure 15 in order to create wire sensor networks.

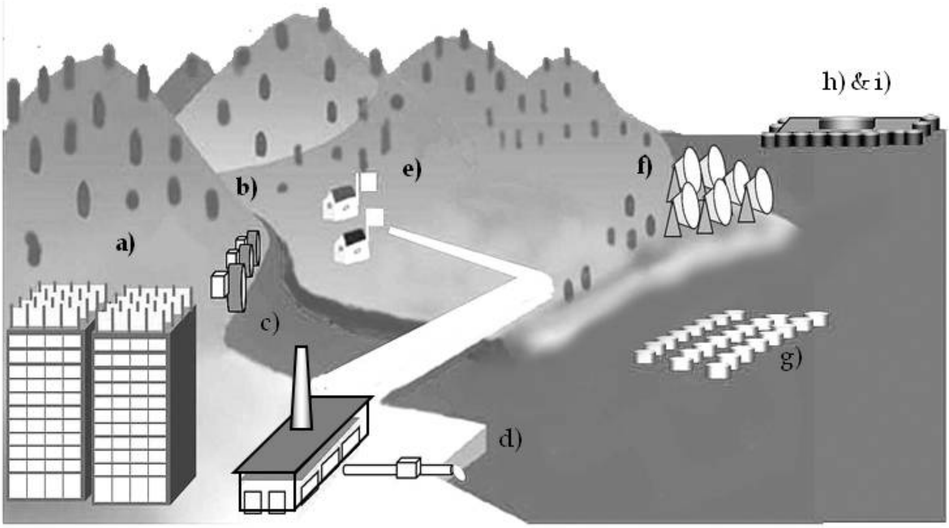


Fig. 15. Sites where power generation using dielectric elastomers is possible and conceptual rendering of the generation systems: (a) Wind Power Generator on tops of buildings (Chiba et al., 2007b) (b) Water Mill Generators (Chiba et al., 2007a) (c) Waste energy Generators (Chiba et al., 2011) (d) Drain Generators (Chiba et al., 2011) (e) Wind Power Generators for Personal Houses (f) Solar Heat Generators (Chiba et al., 2007b) (g) Wave Generators (Chiba et al., 2006; Chiba et al., 2008a) (h) Wave Generators in Ocean (Chiba et al., 2008a) (i) Hydrogen Production Plant (Chiba et al., 2008b)

6. References

Chiba S., et al. (2007a). Extending Applications of Dielectric Elastomer Artificial Muscle. *Proceedings of SPIE*, San Diego, March 2007.

Chiba S.; Stanford S., Pelrine R., Kornbluh R., and Prahlad H. (2006a), Electroactive Polymer Artificial Muscle, JRSJ, Vol. 24, No.4, pp 38-42. 2006.

Chiba S,; Prahad H, Pelrine R, Konbluh R, Stanford S and Eckerle J. (2006b). Electro Power Generation Using Electro active Polymers (EPAM). *Proceedings of 15th Japan Institute of Energy Conference* (Kogakuuin University, Japan) JIE pp 297-298, July 2006.

Chiba S.; Pelrine R., Kornbluh R., Prahlad H., Stanford S., & Eckerle J. (2007b). New Opportunities in Electric Generation Using Electroactive Polymer Artificial Muscle (EPAM). *J. Japan. Inst. Energy*, Vol. 86, No. 9, pp. 38-42, 2007.

Chiba S (2002), Dielectric Elastomer for MEMS and NEMS and Toward the Future. *Electro Packing Technology*, Vol.18, No. 1, pp 33-38, 2002.

Chiba S.; Waki M., Kormbluh R., & Pelrine R. (2008a). Innovative Power Generators for Energy Harvesting Using Electroactive Polymer Artificial Muscles, Electroactive Polymer Actuators and Devices (EAPAD), ed. Y. Bar-Cohen. *Proceedings of SPIE*. Vol. 6927, 692715 (1-9), San Diego, March 2008.

Chiba, S., Kornbluh R., Pelrine R., and Waki M. (2008b) "Low-cost Hydrogen Production From Electroactive Polymer Artificial Muscle Wave Power Generators", *Proceedings of World Hydrogen Energy Conference*, Brisbane, Australia, June 16-20, 2008.

Chiba, S., Waki M., Kornbluh K., and Pelrine R.. (2009). Innovative Wave Power Generation System Using EPAM. *Proceedings of Oceans' 09*, Bremen, Germany, May 2009.

Chiba S. and Waki M. (2011). Recent Progress in Dielectric Elastomers (Harvesting Energy Mode and High Efficient Actuation Mode), To be published in Clean Tech, Nihon Kogyo Shuppan, Tokyo, Japan, April, 2011.

Harsha P, Kornbluh R, Pelrine R, Stanford S, Eckerle J and Oh S. (2005). Polymer Power: Dielectric elastomers and their applications in distributed actuation and power generation. *Proceedings of ISSS 2005, International Conference on Smart Materials Structures and Systems.* Bangalore, India.

Kornbluh R., Pelrine R., and Chiba S. (2004b). Silicon to Siliocon: Stretching the Capabilities of Micromachines with Electroactive polymers, *IEEJ*, Vol.124, No. 8, 2004, ISSN 1341-8939.

Miyazaki T and Osawa H. (2007). Search Report of Wave Power Devices *Proceedings of Spring Conference of the Japan Socity of Naval Architects and Ocean Engineers*, No.4 pp43-46, April 2007.

Pelrine R., and Chiba S. (1992a). Review of Artificial Muscle Approaches. *Proceedings of Third International Symposium on Micromachine and Human Science*, Nagoya, Japan, June 1992.

Pelrine R.; Kornbluh K., Pei Q., & Joseph J. (2000a). High Speed Electrically Actuated Elastomers with Over 100% Strain. *Science* 287: 5454, pp 836–839, 2000.

Pei Q., Rosenthal M., Pelrine R., Stanford S., and Kornbluh R (2003) Multifunctional electroelastmer roll actuators and their application for biomimetic walking robots, proceedings of SPIE, Smart Structures and mterials, Electroactive Polymer Actuators and Devices (EAPAD), ed. Y. Bar-Cohen, San Diego, CA, March 2003.

Stanford S, Bonwit N, Pelrine R, Kornbluh R, Pei Q and Chiba S (2004b). Electro Polymer Artificial Muscle (EPAM) for Biomimetics Robots. *Proceedings of 2nd Conference on Artificial Muscles.* AIST Kansai Center, Osaka, Japan, 2004.

Oguro K., Fujiwara N., Asaka K., Onishi K. and Sewa S. (1999). Polymer electrolyte actuator with gold electrodes. *Proceedings of the SPIE's 6th Annual International Symposium on Smart Structures and Materials, SPIE Proc.* Vol. 3669,(1999), pp. 64-71.

Otero F. and Sansiñena M. (1998). Soft and wet conducting polymers for artificial muscles", *Advanced Materials*, 10 (6), (1998) pp. 491-494.

Osada Y., Okuzaki H. and Hori H. (1992b). A polymer gel with electrically driven motility", *Nature*, Vol. 355, pp. 242-244, (1992).

Ashida A., Ichiki M., Tanaka T. and Kitahara T. (2000b). Power Generation Using Piezo Element: Energy Conversion Efficiency of Piezo Element", *Proc. of JAME annual meeting*, pp.139-140, (2000).