

Communications and Networking

edited by
Jun Peng

SCIYO

Communications and Networking

Edited by Jun Peng

Published by Sciyo

Janeza Trdine 9, 51000 Rijeka, Croatia

Copyright © 2010 Sciyo

All chapters are Open Access articles distributed under the Creative Commons Non Commercial Share Alike Attribution 3.0 license, which permits to copy, distribute, transmit, and adapt the work in any medium, so long as the original work is properly cited. After this work has been published by Sciyo, authors have the right to republish it, in whole or part, in any publication of which they are the author, and to make other personal use of the work. Any republication, referencing or personal use of the work must explicitly identify the original source.

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 Ana Nikolic

Technical Editor Teodora Smiljanic

Cover Designer Martina Sirotic

Image Copyright Alex Staroseltsev, 2010. Used under license from Shutterstock.com

First published September 2010

Printed in India

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

Additional hard copies can be obtained from publication@sciyo.com

Communications and Networking, Edited by Jun Peng

p. cm.

ISBN 978-953-307-114-5

SCIYO.COM
WHERE KNOWLEDGE IS FREE

free online editions of Sciyo
Books, Journals and Videos can
be found at **www.sciyo.com**

Contents

Preface IX

- Chapter 1 **Transform Domain based Channel Estimation for 3GPP/LTE Systems** 1
Moussa Diallo, Rodrigue Rabineau, Laurent Cariou and Maryline H elard
- Chapter 2 **Channel Estimation for Wireless OFDM Communications** 17
Jia-Chin Lin
- Chapter 3 **OFDM Communications with Cooperative Relays** 51
H. Lu, H. Nikookar and T. Xu
- Chapter 4 **High Throughput Transmissions in OFDM based Random Access Wireless Networks** 81
Nuno Souto, Rui Dinis, Jo o Carlos Silva, Paulo Carvalho and Alexandre Louren o
- Chapter 5 **Joint Subcarrier Matching and Power Allocation for OFDM Multihop System** 101
Wenyi Wang and Renbiao Wu
- Chapter 6 **MC-CDMA Systems: a General Framework for Performance Evaluation with Linear Equalization** 127
Barbara M. Masini, Flavio Zabini and Andrea Conti
- Chapter 7 **Wireless Multimedia Communications and Networking Based on JPEG 2000** 149
Max AGUEH
- Chapter 8 **Downlink Capacity of Distributed Antenna Systems in a Multi-Cell Environment** 173
Wei Feng, Yunzhou Li, Shidong Zhou and Jing Wang
- Chapter 9 **Innovative Space-Time-Space Block Code for Next Generation Handheld Systems** 187
Youssef Nasser and Jean-Fran ois H elard
- Chapter 10 **Throughput Optimization for UWB-Based Ad-Hoc Networks** 205
Chuanyun Zou

- Chapter 11 **Outage Probability Analysis of Cooperative Communications over Asymmetric Fading Channel** 221
Sudhan Majhi, Youssef Nasser and Jean François Héland
- Chapter 12 **Indoor Radio Network Optimization** 237
Lajos Nagy
- Chapter 13 **Introduction to Packet Scheduling Algorithms for Communication Networks** 263
Tsung-Yu Tsai, Yao-Liang Chung and Zsehong Tsai
- Chapter 14 **Reliable Data Forwarding in Wireless Sensor Networks: Delay and Energy Trade Off** 289
M. K. Chahine, C. Taddia and G. Mazzini
- Chapter 15 **Cross-Layer Connection Admission Control Policies for Packetized Systems** 305
Wei Sheng and Steven D. Blostein
- Chapter 16 **Advanced Access Schemes for Future Broadband Wireless Networks** 323
Gueguen Cédric and Baey Sébastien
- Chapter 17 **Medium Access Control in Distributed Wireless Networks** 339
Jun Peng
- Chapter 18 **Secure Trust-based Cooperative Communications in Wireless Multi-hop Networks** 359
Kun Wang, Meng Wu and Subin Shen
- Chapter 19 **Wireless Technologies and Business Models for Municipal Wireless Networks** 379
Zhe Yang and Abbas Mohammed
- Chapter 20 **Data-Processing and Optimization Methods for Localization-Tracking Systems** 389
Giuseppe Destino, Davide Macagnano and Giuseppe Abreu
- Chapter 21 **Usage of Mesh Networking in a Continuous-Global Positioning System Array for Tectonic Monitoring** 415
Hoang-Ha Tran and Kai-Juan Wong

Preface

This book “Communications and Networking” focuses on the issues at the lowest two layers of communications and networking and provides recent research results on some of these issues. In particular, it first introduces recent research results on many important issues at the physical layer and data link layer of communications and networking and then briefly shows some results on some other important topics such as security and the application of wireless networks.

This book has twenty one chapters that are authored by researchers across the world. Each chapter introduces not only a basic problem in communications and networking but also describes approaches to the problem. The data in most chapters are based on published research results and provide insights on the problems of the relevant chapters. Most chapters in this book also provide references for the relevant topics and interested readers might find these references useful if they would like to explore more on these topics.

Several chapters of this book focus on issues related to Orthogonal Frequency-Division Multiplexing (OFDM). For example, chapter 1 and chapter 2 are on channel estimation for OFDM-related systems. Chapter 3 is on cooperative relays in OFDM systems. Chapter 4 introduces some recent results on packet separation in OFDM based random access wireless networks. Chapter 4 is on sub-carrier matching and power allocation in OFDM-based multihop systems. Chapter 6 presents some results on performance evaluation of OFDM related systems.

Multiple chapters of this book are on coding, link capacity, throughput, and optimisation. For example, chapter 7 and chapter 9 are about source and channel coding in communications and networking. Chapter 8 is on link capacity in distributed antenna systems. Chapter 10 introduces throughput optimisation for UWB-based ad hoc networks. Chapter 12 presents some results on optimising radio networks.

This book also contains several chapter on forwarding, scheduling, and medium access control in communications and networking. In particular, chapter 13 introduces packet scheduling algorithms for communication networks. Chapter 14 is about reliable data forwarding in wireless sensor networks. Chapter 15 introduces cross-layer connection admission control in packetized systems. Chapter 16 presents advanced access schemes for future broadband wireless networks. Chapter 17 introduces medium access control in distributed wireless networks. Finally, chapter 18 is about cognitive radio networks.

In addition, this book has several chapters on some other issues of communications and networking. For example, chapter 19 is about security of wireless LANs and wireless multihop networks, chapter 20 is on localisation and tracking and chapter 21 introduces the use of mesh networks in tectonic monitoring.

In summary, this book covers a wide range of interesting topics of communications and networking. The introductions, data, and references in this book would help the readers know more about communications and networking and help them explore this exciting and fact-evolving field.

Editor

Jun Peng

*University of Texas - Pan American,
Edinburg, Texas,
United States of America*

Transform Domain based Channel Estimation for 3GPP/LTE Systems

Moussa Diallo¹, Rodrigue Rabineau¹, Laurent Cariou¹ and Maryline Hélard²

¹Orange Labs, 4 rue du Clos Courtel, 35512 Cesson-Sévigné Cedex,

²INSA Rennes, 20 Avenue des Buttes de Coesmes, 35700 Rennes Cedex
France

1. Introduction

Orthogonal frequency division multiplexing (OFDM) is now well known as a powerful modulation scheme for high data rate wireless communications owing to its many advantages, notably its high spectral efficiency, mitigation of intersymbol interference (ISI), robustness to frequency selective fading environment, as well as the feasibility of low cost transceivers [1].

On the other hand multiple input multiple output (MIMO) systems can also be efficiently used in order to increase diversity and improve performance of wireless systems [2] [3] [4]. Moreover, as OFDM allows a frequency selective channel to be considered as flat on each subcarrier, MIMO and OFDM techniques can be well combined. Therefore, MIMO-OFDM systems are now largely considered in the new generation of standards for wireless transmissions, such as 3GPP/LTE [5] [6].

In most MIMO-OFDM systems, channel estimation is required at the receiver side for all sub-carriers between each antenna link. Moreover, since radio channels are frequency selective and time-dependent channels, a dynamic channel estimation becomes necessary. For coherent MIMO-OFDM systems, channel estimation relies on training sequences adapted to the MIMO configuration and the channel characteristics [7] and based on OFDM channel estimation with pilot insertion, for which different techniques can be applied: preamble method and comb-type pilot method.

In order to estimate the channel of an OFDM systems, one's first apply least square (LS) algorithm to estimate the channel on the pilot tones in the frequency domain. A second step can be performed to improve the quality of the estimation and provide interpolation to find estimates on all subcarriers. In a classical way, this second step is performed in the frequency domain. An alternative is to perform this second step by applying treatment in a transform domain, that can be reached after a discrete Fourier transform (DFT) or a discrete cosine transform (DCT), and called transform domain channel estimation (TD-CE). The DFT based method is considered as a promising method because it can provide very good results by significantly reducing the noise on the estimated channel coefficients [8]. However, some performance degradations may occur when the number of OFDM inverse fast fourier transform (IFFT) size is different from the number of modulated subcarriers [8]. This problem called "border effect" phenomenon is due to the insertion of null carriers at the spectrum extremities (virtual carriers) to limit interference with the adjacent channels, and can be encountered in most of multicarrier systems.

To cope for this problem, DCT has been proposed instead of DFT, for its capacity to reduce the high frequency components in the transform domain [9]. Its improvements are however not sufficient in systems designed with a great amount of virtual subcarriers, which suffer from a huge border effect [10]. This is the case of a 3GPP/LTE system.

The aim of the paper is to study, for a 3GPP/LTE system, two improved DCT based channel estimations, designed to correctly solve the problem of null carriers at the border of the spectrum. These two TD-CE will also be compared in terms of performance and complexity. In the first approach, a truncated singular value decomposition (TSVD) of pilots matrix is used to mitigate the impact of the "border effect". The second approach is based on the division of the whole DCT window into 2 overlapping blocks.

The paper is organized as follows. Section II introduces the mobile wireless channel and briefly describes the MIMO-OFDM system with channel estimation component. Section III is dedicated to transform domain channel estimations (TD-CE), with description of the classical Least Square algorithm in III-A, and presents the conventional DFT and DCT based channel estimation in III-B and III-C, respectively. Next, the two proposed DCT based channel estimation are described in the sections IV and V. Finally, a performance evaluation and comparison is shown in section VI.

2. MIMO-OFDM system description

In this paper we consider a coherent MIMO-OFDM system, with N_t transmit antennas and N_r receive antennas. As shown in Fig.1, the MIMO scheme is first applied on data modulation symbols (e.g. PSK or QAM), then an OFDM modulation is performed per transmit antenna. Channel estimation is then required at receive side for both the one tap per sub-carrier equalization and the MIMO detection.

The OFDM signal transmitted from the i -th antenna after performing IFFT (OFDM modulation) to the frequency domain signal $X_i \in \mathbb{C}^{N \times 1}$ can be given by:

$$x_i(n) = \sqrt{\frac{1}{N}} \sum_{k=0}^{N-1} X_i(k) e^{j \frac{2\pi kn}{N}}, \quad 0 \leq (n, k) \leq N \quad (1)$$

where N is the number of FFT points.

The time domain channel response between the transmitting antenna i and the receiving antenna j under the multipath fading environments can be expressed by the following equation:

$$h_{ij}(n) = \sum_{l=0}^{L-1} h_{ij,l} \delta(n - \tau_{ij,l}) \quad (2)$$

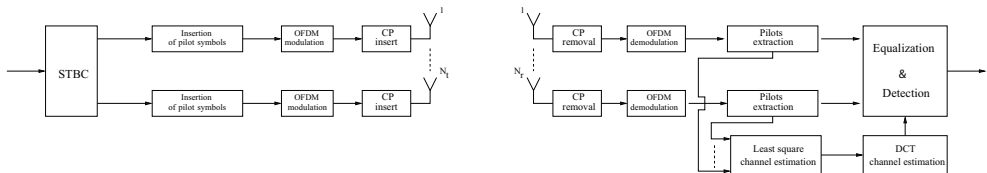


Fig. 1. MIMO-OFDM block diagram.

with L the number of paths, $h_{ij,l}$ and $\tau_{ij,l}$ the complex time varying channel coefficient and delay of the l -th path.

The use of a guard interval allows both the preservation of the orthogonality between the tones and the elimination of the inter symbol interference (ISI) between consecutive OFDM symbols. Thus by using (1) and (2), the received frequency domain signal is given by:

$$R_j(k) = \sum_{i=0}^{N_t-1} X_i(k)H_{ij}(k) + \Xi(k) \quad (3)$$

where $H_{ij}(k)$ is the discrete response of the channel on subcarrier k between the i -th transmit antenna and the j -th receive antenna and Ξ_k the zero-mean complex Gaussian noise after the FFT (OFDM demodulation) process.

3. Transform Domain Channel Estimation (TD-CE)

In a classical coherent SISO-OFDM system, channel estimation is required for OFDM demodulation. When no knowledge of the statistics on the channel is available, a least square (LS) algorithm can be used in order to estimate the frequency response on the known pilots that had been inserted in the transmit frame. An interpolation process allows then the estimation of the frequency response of the channel, i.e. for each sub-carrier. In a MIMO-OFDM system, since the received signal is a superposition of the transmitted signals, orthogonally between pilots is mandatory to get the channel estimation without co-antenna interference (CAI).

We choose to apply TD-CE to a 3GPP/LTE system where the orthogonality between training sequences is based on the simultaneous transmission on each subcarrier of pilot symbols on one antenna and null symbols on the other antennas as depicted in Fig.2.

A. Least Square channel estimation (LS)

Assuming orthogonality between pilots dedicated to each transmit antenna, the LS estimates can be expressed as follows:

$$H_{ij,LS} = H_{ij} + (\text{diag}(X))^{-1}\Xi. \quad (4)$$

Therefore LS estimates can be only calculated for $\frac{M}{N_t}$ subcarriers where M is the number of modulated subcarriers. Then interpolation has to be performed to obtain an estimation for all the subcarriers.

B. DFT based channel estimation

From (4), it can be observed that LS estimates can be strongly affected by a noise component. To improve the accuracy of the channel estimation, the DFT-based method has been proposed in order to reduce the noise component in the time domain [8]. Fig.3 illustrates the transform domain channel estimation process using DFT. After removing the unused subcarriers, the LS estimates are first converted into the time domain by the IDFT algorithm and a smoothing filter (as described in Fig.3) is applied in the time domain assuming that the maximum multi-path delay is within the cyclic prefix of the OFDM symbols. After the smoothing, the DFT is applied to return in the frequency domain.

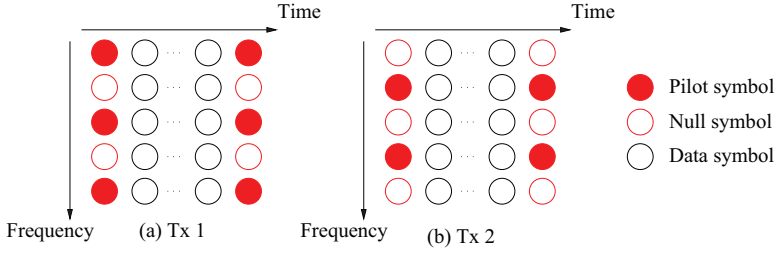


Fig. 2. Pilot insertion structure in 3GPP with $N_t = 2$.

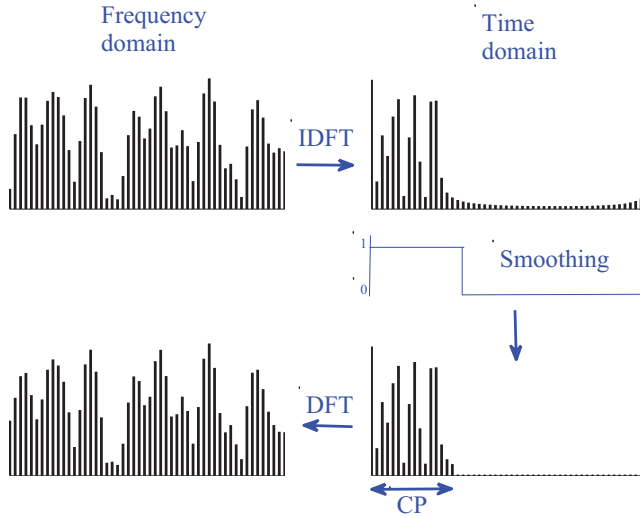


Fig. 3. Transform domain channel estimation process using DFT.

The time domain channel response of the LS estimated channel can be expressed by (5). From (4), it is possible to divide $h_{n,LS}^{IDFT}$ into two parts.

$$\begin{aligned} h_{n,LS}^{IDFT} &= \sqrt{\frac{1}{M-1}} \sum_{k=0}^{M-1} H_{k,LS} e^{j\frac{2\pi nk}{M}} \\ &= h_n^{IDFT} + \xi_n^{IDFT} \end{aligned} \quad (5)$$

where ξ_n^{IDFT} is the noise component in the time domain and h_n^{IDFT} is the IDFT of the LS estimated channel without noise which is developed as:

$$h_n^{IDFT} = \sqrt{\frac{1}{M}} \sum_{l=0}^{L-1} h_l e^{-j\pi\tau_l(1-\frac{M}{N})} \sum_{k=0}^{M-1} e^{-j\frac{2\pi k}{M}(\frac{M}{N}\tau_l - n)} \quad (6)$$

It can be easily seen from (6) that if the number of FFT points N is equal to the number of modulated subcarriers M , the impulse response h_n^{IDFT} will exist only from $n = 0$ to $L - 1$, with the same form as (2), i.e the true channel.

Nevertheless when $N > M$, the last term of (6) $\sum_{k=0}^{M-1} e^{-j\frac{2\pi k}{M}(\frac{M}{N}\tau_l - n)}$ can be expressed as.

$$\begin{cases} M & \frac{M / hcf(M, N)}{N / hcf(M, N)} \tau_l : \leq L \text{ and } \in \mathbb{N} \\ \frac{1 - e^{-j2\pi(\frac{M}{N}\tau_l - n)}}{1 - e^{-j\frac{2\pi}{M}(\frac{M}{N}\tau_l - n)}} & \text{otherwise} \end{cases} \quad (7)$$

where hcf is the highest common factor and \mathbb{N} natural integer.

From (7) it is important to note that:

- On the one hand, the channel taps are not all completely retrieved in the first CP samples of the channel impulse response.
- On the other hand, the impulse channel response obviously exceeds the Guard Interval (CP). This phenomenon is called Inter-Taps Interference (ITI). Removing the ITI by the smoothing filter generates the “border effect” phenomenon.

C. DCT based channel estimation

The DCT based channel estimator can be realized by replacing IDFT and DFT (as shown in Fig.3) by DCT and IDCT, respectively. DCT conceptually extends the original M points sequence to $2M$ points sequence by a mirror extension of the M points sequence [12]. As illustrated by Fig. 4, the waveform will be smoother and more continuous in the boundary between consecutive periods.

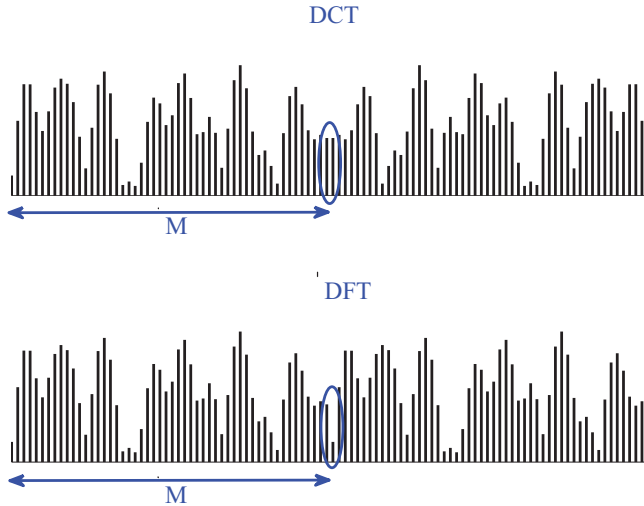


Fig. 4. DCT and DFT principle.

The channel impulse response in the transform domain is given by:

$$h_{n,LS}^{DCT} = V_n^M \sum_{k=0}^{M-1} H_{k,LS} \cdot \cos\left(\frac{\pi(2k+1)n}{2M}\right) \quad (8)$$

where V_n^M is the coefficient of DCT which can take two different values, depending on the value of n .

$$V_n^M = \begin{cases} \sqrt{1/M} & n = 0 \\ \sqrt{2/M} & n \neq 0 \end{cases} \quad (9)$$

From the DCT calculation and the multi-path channel characteristics, the impulse response given by (8) is concentrated at lower order components in the transform domain. It is important to note that the level of impulse response at the order higher than N_{max} is not null, but can be considered as negligible; this constitutes the great interest of the DCT. The channel response in the transform domain can be expressed by:

$$h_n^{DCT} = \begin{cases} h_{n,LS}^{DCT} & 0 \leq n \leq N_{max} - 1 \\ 0 & N_{max} \leq n \leq M - 1 \end{cases} \quad (10)$$

The frequency channel response is then given by:

$$H_k^{DCT} = \sum_{n=0}^{M-1} V_n^M h_n^{DCT} \cdot \cos\left(\frac{\pi(2k+1)n}{2M}\right) \quad (11)$$

As a summary of this conventional DCT based estimation, it is important to note this following remark:

In the conventional DCT based method, the ITI is less important than in DFT one but a residual "border effect" is still present.

4. DCT with TSVD based channel estimation

In the classical DCT approach, it is shown that all the channel paths are retrieved. Nevertheless, the residual ITI will cause the "border effect". The following approach is a mixture of Zero Forcing (ZF) and a truncated singular value decomposition in order to reduce the impact of null subcarriers in the spectrum [13]. The DCT transfer matrix C of size $N \times N$ can be defined with the following expression:

$$C = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & D_N & \dots & D_N(2N+1) \\ \vdots & \vdots & \dots & \vdots \\ 1 & D_N(N-1) & \dots & D_N((2N+1)(N-1)) \end{bmatrix} \quad (12)$$

where $D_N(kn) = V_n^N \cdot \cos\left(\frac{\pi}{2N}(2k+1)(n)\right)$.

To accommodate the non-modulated carriers, it is necessary to remove the rows of the matrix C corresponding to the position of null subcarriers (see Fig.2). From (10), we can just

use the first N_{max} columns of C . Hence the DCT transfer matrix becomes:

$$\tilde{C}'_i = C\left(\frac{N-M}{2} + i : N_t : \frac{N+M}{2} - 1, 1 : N_{max}\right) \text{ where } 0 \leq i \leq N_t \text{ is the transmit antenna index.}$$

Let us rewrite (8) in a matrix form:

$$h_{LS}^{DCT} = \tilde{C}' \cdot H_{LS} \quad (13)$$

To mitigate the ITI, the first step of this new approach is to apply the ZF criterion [14]:

$$h_{LS}^{IDCT-ZF} = (\tilde{C}'^H \tilde{C}')^{-1} \tilde{C}'^H H_{LS} = \tilde{C}'^\dagger H_{LS} \quad (14)$$

The main problem arises when the condition number (CN) of $\tilde{C}'^H \tilde{C}'$, defined by the ratio between the greater and the lower singular value, becomes high. Fig.5 shows the behavior of the singular value of $\tilde{C}'^H \tilde{C}'$ whether null carriers are placed at the edge of the spectrum or not. When all the subcarriers are modulated, the singular values are all the same and the CN is equal to 1. However, when null carriers are placed at the edges of the spectrum, the CN becomes very high. For instance, as we can see in Fig.5, if $N = 1024$, $N_{max} = 84$ and $M = 600$ as in 3GPP, the CN is 2.66×10^{16} .

To reduce the "border effect", i.e the impact of ITI, it is necessary to have a small condition number. The second step of this new approach is to consider the truncated singular value decomposition (TSVD) of the matrix $\tilde{C}'^H \tilde{C}'$ of rank N_{max} .

Fig.6 shows the block diagram of the DCT based channel estimation and the proposed scheme. In the proposed scheme (Fig.6(b)), after performing the SVD of the matrix $\tilde{C}'^H \tilde{C}'$, we propose to only consider the T_h most important singular values among the N_{max} in order to reduce the CN. The TSVD solution is defined by:

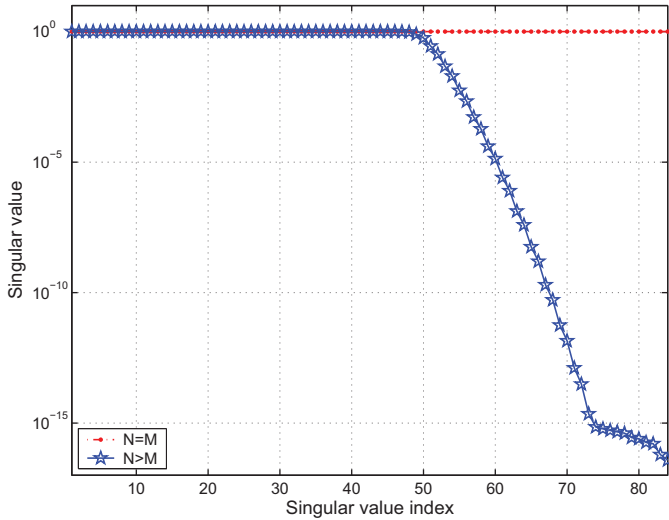


Fig. 5. Singular value of $\tilde{C}'^H \tilde{C}'$ with $N_{max} = 84$, $CP = 72$ and $N = M = 1024$ or $N = 1024$, $M = 600$

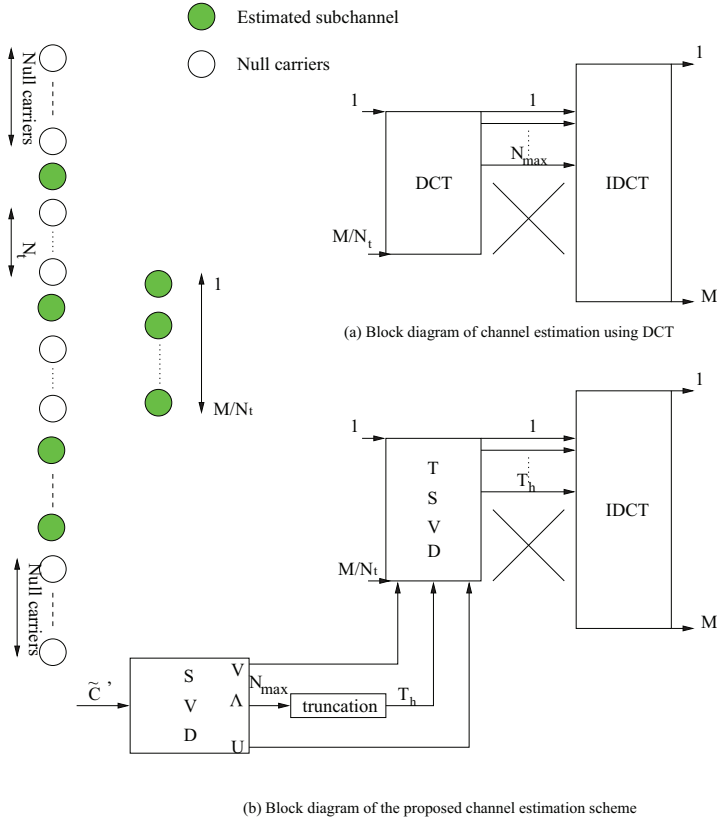


Fig. 6. Block diagram of channel estimation using DCT and the proposed scheme.

$$H_{n,LS}^{DCT-ZF-TSVD} = \sum_{s=1}^{T_h} \frac{u_s^H H_{n,LS} v_s}{\sigma_s} \quad (15)$$

where T_h is the threshold, u_s , v_s and σ_s are the left singular vector, the right singular vector and the singular values of \tilde{C}' .

An IDCT (\tilde{C}'^H) is then used to get back to the frequency domain.

$$H_k^{DCT-ZF-TSVD} = C^{global} = \tilde{C}'^H \tilde{C}' \quad (16)$$

$T_h (\in 1, 2, \dots, N_{max})$ can be viewed as a compromise between the accuracy on pseudo-inverse calculation and the CN reduction. The adjustment of T_h is primarily to enhance the channel estimation quality. Its value depends only on the system parameters (position of the null carriers), which is predefined and known at the receiver side. T_h can be in consequence calculated in advance for any MIMO-OFDM system. To find a good value of T_h , is important to master its effect on the channel estimation i.e on the matrix $C^{global} \in \mathbb{C}^{M/N_t \times M}$. As an example, Fig.7 shows the behavior of the M/N_t singular values of C^{global} for different T_h where CP = 72, N = 1024, M = 600 and $N_t = 4$.

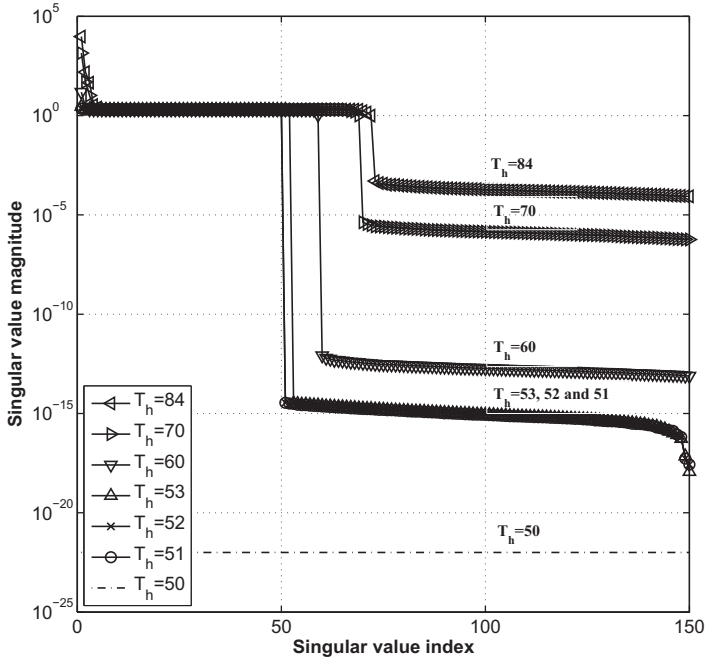


Fig. 7. Singular values of $\tilde{C}^H \tilde{C}^T$ with $CP = 72$, $N = 1024$ and $M = 600$ for different values of T_h

For $T_h = 51, 52, 53$ the singular values of C^{global} are the same on the first T_h samples and almost zero for others samples. We can consider that the rank of the matrix C^{global} becomes T_h instead of N_{max} . Therefore the noise effect is minimized and CN is equal to 1.

However, all the singular values become null when $T_h = 50$ due to a very large loss of energy. As illustrated by the Fig.8 which is a zoom of Fig.7 on the first singular values, their behavior can not be considered as a constant for $T_h = 60, 70, 84$ and then the CN becomes higher.

5. DCT with 2 overlapping blocks

The principle of this approach is to divide the whole DCT window into R blocks as proposed in [18]. In this paper we consider $R = 2$, that was demonstrated to reach same bit error rate (BER) performance that higher R values.

As illustrated in Fig.9, the concatenation of the 2 overlapping blocks cannot exceed N .

The classical DCT smoothing process described in the section III-C is applied to each 2 blocks of size $N/2$ by keeping only the energy of the channel in the first $N_{max}/2$ samples. However, the residual ITI causes "border effect" on the edge of each block. Then, to recover the channel coefficients, we average the values in the overlapping windows between the different blocks except some subcarriers at the right and the left edge of block 1 and block 2 respectively as described in Fig.10.

The noise power is averaged on N samples instead of M in this approach. Thereby it presents a gain $(10 \log_{10}(\frac{N}{M}))$ in comparison to the classical DCT based channel estimation.

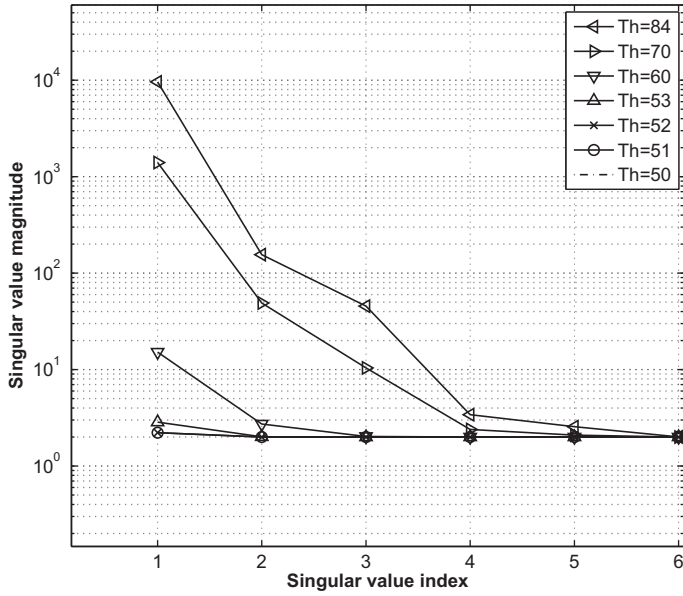


Fig. 8. Singular values of $\tilde{C}^H \tilde{C}^\dagger$ with $CP = 72$, $N = 1024$, $M = 600$ and $Nt = 4$ for different values of T_h

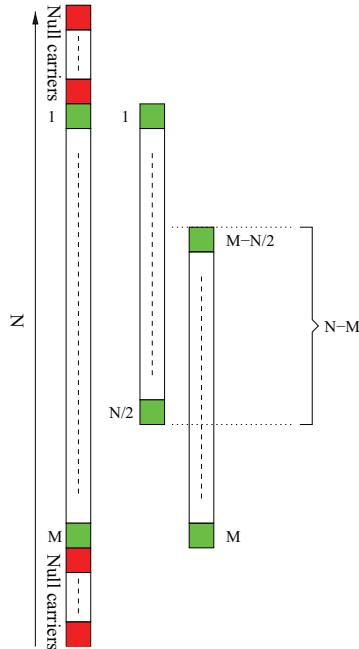


Fig. 9. Principle of the DCT with 2 overlapping blocks

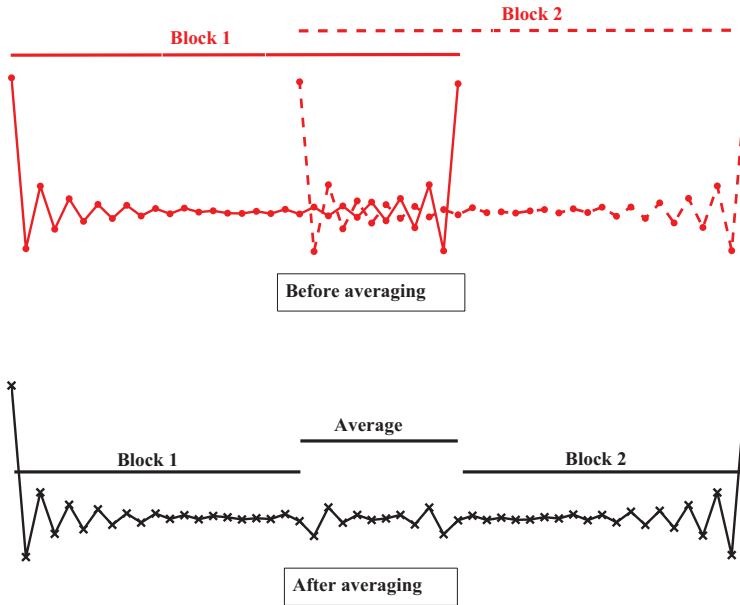


Fig. 10. Recovery of the channel coefficients.

For instance, the gain is 2.31dB for the studied 3GPP/LTE system ($N = 1024$ $M = 600$).

6. Simulations results

The different channel estimation techniques, LS estimation, classical DFT and DCT estimations and the two proposed DCT estimations (DCT with TSVD and DCT with 2 overlapping blocks) are applied to a 4×2 MIMO-OFDM system with a double-Alamouti scheme. After the description of the system parameters, the performance and complexity of the channel estimation techniques will be analysis. Note that DCT-TSVD method is named on the figures by the used threshold (DCT-TSVD with $T_h = 53$ is named $T_h = 53$), while DCT with 2 overlapping blocks is called DCT_2 .

A. System parameters

Performance are provided over frequency and time selective MIMO SCME typical to urban macro channel model (C) without any spatial correlation between transmit antennas [15]. Double- Alamouti space-time coding consists in simultaneously transmitting two Alamouti codes on two blocks of two transmit antennas [16].

The system parameters are issued and close to those defined in 3GPP/LTE framework [6]. The detailed parameters of the system simulations are listed in Table I.

B. Performances analysis

Fig.11 shows mean square error (MSE) on different subcarriers for the proposed DCT-TSVD based channel estimation with the optimized threshold $T_h = 53$, the proposed DCT with 2 overlapping blocks and the conventional DFT and DCT ones in 3GPP/LTE system. We can

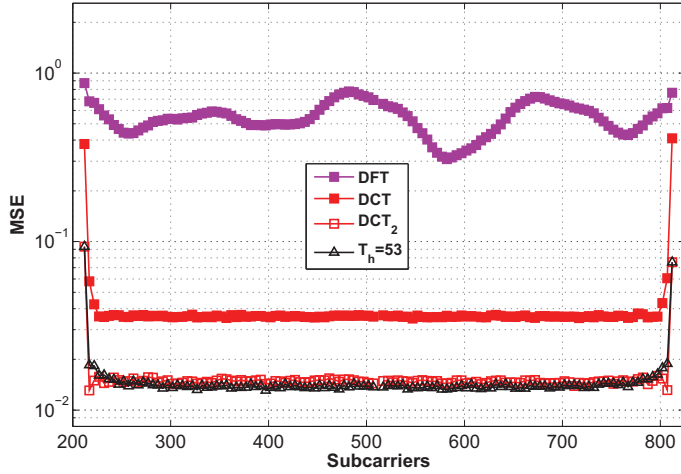


Fig. 11. MSE per subcarriers for 3GPP/LTE: $E_b/N_0 = 10dB$.

Channel Model	SCME Channel Model C
Number of FFT points (N) & Modulated subcarriers (M)	1024 & 600
cyclic prefix	72
Number of N_t & N_r antennas	4 & 2
Bandwidth & Carrier frequency	15.36MHz & 2GHz
Modulation & Coding Rate	16QAM & 1/3
MIMO scheme	double-Alamouti
FEC	turbo code (UMTS)

Table I. Simulation parameters

first see that DCT based channel estimation reduces significantly the “border effect” in comparison to the conventional DFT one. The two proposed optimized DCT methods allow MSE to be improved on all subcarriers even at the edges of the spectrum compared to the conventional DCT one. For DCT with 2 overlapping blocks, this can be explained by the noise reduction obtained thanks to the averaging which is performed on the overlapped portion of the spectrum. For DCT-TSVD method, improvement is due to the minimization of the noise effect and the reduction of the CN obtained by using TSVD calculation. The MSE performance, averaged over all subcarriers, can be observed in Fig.12 which shows MSE versus E_b/N_0 , for the different channel estimation techniques. Note that the two optimized techniques, DCT-TSVD and DCT with 2 overlapping blocks, present very similar performance.

This can also be observed in Fig.13, which represents the performance results in terms of BER versus E_b/N_0 for perfect, least square (LS), classical DFT and DCT, the proposed DCT-TSVD channel estimation with $T_h = 84, 70, 60, 53, 52, 51$ and the proposed DCT with 2 overlapping blocks. The classical DFT based method presents poor results due to the

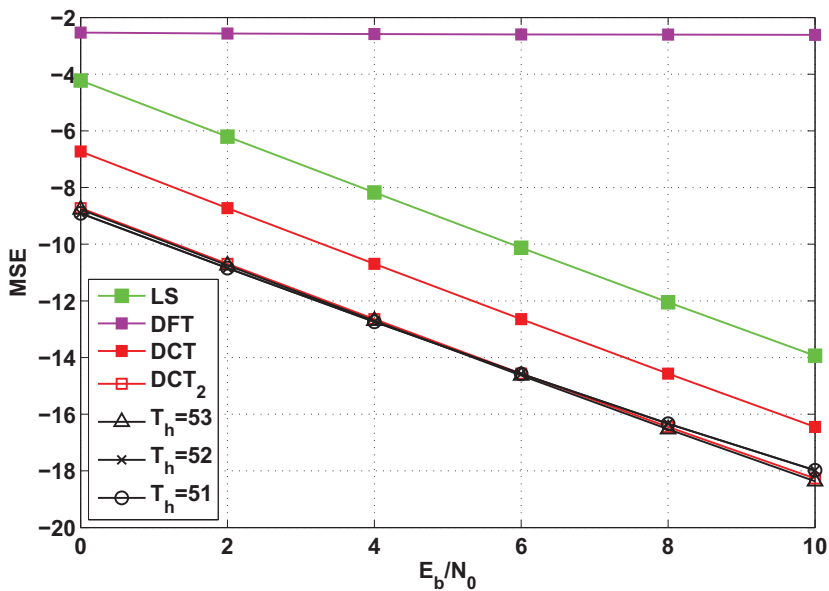


Fig. 12. MSE against E_b/N_0 for 3GPP/LTE.

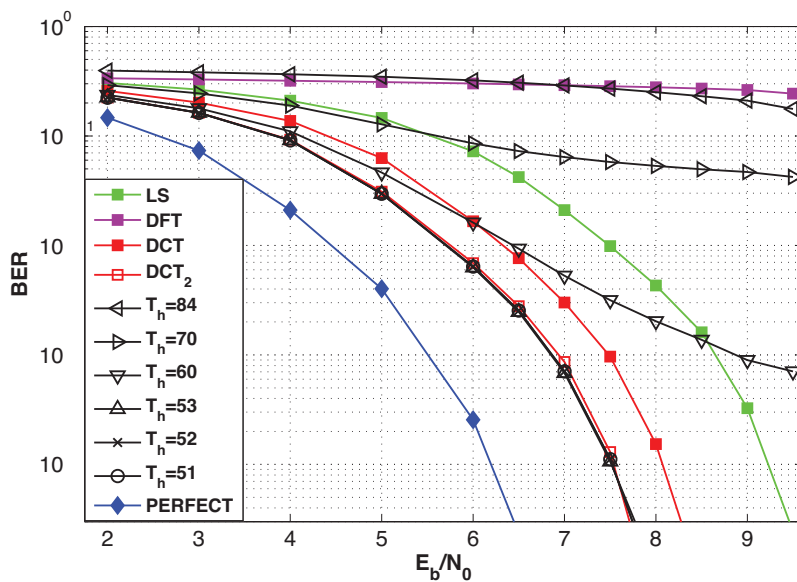


Fig. 13. BER against E_b/N_0 for 3GPP/LTE.

“border effect” whereas the DCT one improves the accuracy of the channel estimation by significantly reducing the noise component compared to the LS estimate (1.5 dB).

The DCT-TSVD estimator with $T_h = 84, 70, 60$ presents an error floor due the high CN. The optimization of T_h ($T_h = 53, 52, 51$) allows the residual “border effect” to be mitigated and then the noise component to be reduced. The performance is then improved compared to the classical DCT. The proposed DCT with 2 overlapping blocks also improves the performance compared to the classical DCT due to the large noise reduction.

It is important to note that, the noise reduction gain obtained by using DCT with 2 overlapping blocks depends on the system parameters ($10\log_{10}(\frac{N}{M})$). In 3GPP/LTE context ($N = 1024$ and $M = 600$), the gain is very important (2.3dB), which allows DCT with 2 overlapping blocks to have the same performance than the optimized DCT-TSVD.

C. Complexity analysis

Considering a MIMO-OFDM system with N_t transmit antennas and N_r received ones, the channel estimation module is used $N_t \times N_r$ times to estimate the MIMO channel between all the antenna links. Thus, the complexity of the channel estimation can be very high in term of real multiplications and real additions.

- In DCT-TSVD, the global matrix C^{global} will be used $N_t \times N_r$ times to estimate the MIMO channel. The number of real multiplications and real additions are $(2M^2/N_t)N_tN_r$ and $(2M(M/N_t - 1))N_tN_r$, respectively.
- In order to use the inverse fast fourier transform (IFFT) and fast fourier transform (FFT) algorithms for OFDM modulation and demodulation, the number of subcarriers (N) is chosen as a power of 2 in all multicarriers systems. For instance, $N = 2^6 = 64$ for IEEE802.11n, 2^{10} for 3GPP/LTE and 2^{11} for digital video broadcasting terrestrial DVBT 2k. The proposed DCT with 2 overlapping blocks uses two blocks of size $N/2$ which is a power of 2. Therefore, this channel estimation technique can be performed by using fast DCT algorithms [17]. Then, the number of real multiplications and real additions are $(\frac{N}{4})\log_2(\frac{N}{2}) \times 4N_tN_r$ and $((\frac{3N}{4})\log_2(\frac{N}{2}) - \frac{N}{2} + 1) \times 4N_tN_r$, respectively.

Algorithm	Multiplications	Additions
DCT with TSVD	1440000	1431040
DCT with 2 overlapping blocks	73728	204832

Table II. Number of required operations to estimate the $N_t \times N_r$ subchannels ($N_t = 4, N_r = 2$)

Table II contains the number of required operations to estimate the 4×2 subchannels, for a 3GPP-LTE system. We can clearly see in this table that the complexity reduction with 2 overlapping blocks, compared to DCT-TSVD, is very important. As in the context of the study (3GPP/LTE parameters), the performance results of the two techniques are very close, the DCT channel estimation with 2 overlapping blocks becomes a very interesting and promising solution.

7. Conclusion

In this paper, two improved DCT based channel estimations are proposed and evaluated in a 3GPP/LTE system context. The first technique is based on truncated singular value

decomposition (TSVD) of the transfer matrix, which allows the reduction of the condition number (CN). The second technique is based on the division of the whole DCT window into 2 overlapping blocks. The noise reduction gain obtained by using DCT with 2 overlapping blocks, which depends on the system parameters, is very important.

The simulation results in 3GPP/LTE context show that the performance results of the two techniques are very close but the DCT channel estimation with 2 overlapping blocks becomes a very interesting and promising solution due to its low complexity. It can be noted that this complexity could be further reduced by considering more than two blocks.

8. References

- [1] S. B. Weinstein and P.M. Ebert. Data transmission by frequency-division multiplexing using Discret Fourier Transform. *IEEE Trans.Commun.*, Vol. COMM-19, pp.628-634, Oct. 1971.
- [2] I. E. Telatar. Capacity of Multi-antenna Gaussian Channel . ATT Bell Labs tech.memo, June 1995.
- [3] S. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE J. Select. Areas Communication*, Vol. 16, pp. 1451-1458, Oct. 1998.
- [4] V. Tarokh, H. Jafarkhani, and A. R. Calderbank. Space-time block codes from orthogonal designs. *IEEE trans. Inform. Theory*, Vol.45, pp.1456-1467, July 1999.
- [5] N. Boubaker, K.B. Letaief, and R.D. Murch. A low complexity multi-carrier BLAST architecture for realizing high data rates over dispersive fading channels . *IEEE VTC*, Vol. 2, May 2001.
- [6] 3GPP TSG-RAN. 3gpp tr 25.814. physical layer aspects for evolved ultra (release 7).. doc.: Technical report, 2006.
- [7] J. Moon, H. Jin, T. Jeon, and S.-K. Lee. Channel Estimation for MIMO-OFDM Systems Employing Spatial Multiplexing. doc.: *IEEE VTC*, volume5, pages 36-49, 2004.
- [8] E.G. Larsson, and J. Li. Preamble Design for Multiple-Antenna OFDM-based WLANs with Null Subcarriers. *IEEE Signal Processing Letters*, Vol. 8, pp. 285-288, Nov. 2001.
- [9] H. Kobayaki and K. Mori. Proposal of OFDM channel estimation method using discrete cosine transform. *IEEE*, 2004.
- [10] D. Moussa, L. Boher, R. Rabineau, L. Cariou and M. H'elard. Transform Domain Channel Estimation with null Subcarriers for MIMO-OFDM systems. *IEEE trans*, ISWCS 2008.
- [11] D.S.Baum, J.Hansen, G.Del Galdo, M. Milojevic, J. Salo and P. Kyösti. An interim channel model for beyond-3g systems: extending the 3gpp spatial channel model (smc).. doc.: *IEEE VTC*, volume5, pages 3132-3136, May 2005.
- [12] E. Rivier. *Communication Audiovisuelle*. Springer, pp.395-396, Dec. 2003.
- [13] D. Moussa, R. Rabineau and L. Cariou. Robust DCT based Channel Estimation for MIMO-OFDM systems. *WCNC 2009*.
- [14] X.G. Doukopoulos, and R. Legouable. Robust Channel Estimation via FFT Interpolation for Multicarrier Systems. *IEEE Transactions on Signal Processing*, VTC2007-Spring , 2007.
- [15] D.S.Baum, J.Hansen, G.Del Galdo, M. Milojevic, J. Salo and P. Kyösti. An interim channel model for beyond-3g systems: extending the 3gpp spatial channel model

- (smc).. doc.: IEEE Vehicular Technology Conference, volume5, pages 1067-1071, Nov. 1998.
- [16] S. Baro, G. Bauch, A. Pavlic, and A. Semmler. Improving blast performance using space-time block codes and turbo-decoding.. doc.: IEEE GLOBECOM Conference, volume5, pages 3132-3136, May 2005.
- [17] B.L.Lee A new algorithm to compute the discrete cosine transform. IEEE, on A.S.S0P., Vol.32 N₀.6 1984.
- [18] M. Diallo, R. Rabineau, L. Cariou and M. Helard. On improved DCT based Channel Estimation with very low complexity for MIMO-OFDM systems. VTC spring,pp. 1-5, Barcelona (Spain), April 2009.

Channel Estimation for Wireless OFDM Communications

Jia-Chin Lin
National Central University
Taiwan

1. Introduction

1.1 Preliminary

Orthogonal frequency-division multiplexing (OFDM) communication techniques have recently received significant research attention because of their ability to maintain effective transmission and highly efficient bandwidth utilization in the presence of various channel impairments, such as severely frequency-selective channel fades caused by long multipath delay spreads and impulsive noise (Bingham, 1990; Zou & Wu, 1995). In an OFDM system, a high-rate serial information-bearing symbol stream is split into many low-rate parallel streams; each of these streams individually modulates a mutually orthogonal sub-carrier. The spectrum of an individual sub-channel overlaps with those expanded from the adjacent sub-channels. However, the OFDM sub-carriers are orthogonal as long as they are synthesized such that the frequency separation between any two adjacent sub-carriers is exactly equal to the reciprocal of an OFDM block duration. A discrete Fourier transform (DFT) operation can perfectly produce this sub-carrier arrangement and its relevant modulations (Darlington, 1970; Weinstein & Ebert, 1971). Because of the advanced technologies incorporated into integrated circuit (IC) chips and digital signal processors (DSPs), OFDM has become a practical way to implement very effective modulation techniques for various applications. As a result, OFDM technologies have recently been chosen as candidates for 4th-generation (4G) mobile communications in a variety of standards, such as IEEE 802.16 (Marks, 2008) and IEEE 802.20 (Klerer, 2005) in the United States, and international research projects, such as EU-IST-MATRICE (MATRICE, 2005) and EU-IST-4MORE (4MORE, 2005) for 4G mobile communication standardization in Europe. Regarding the history of OFDM, recall that Chang published a paper on the synthesis of band-limited signals for parallel multi-channel transmission in the 1960s (Chang, 1966). The author investigated a technique for transmitting and receiving (transceiving) parallel information through a linear band-limited channel without inter-channel interference (ICI) or inter-symbol interference (ISI). Saltzberg then conducted relevant performance evaluations and analyses (Saltzberg, 1967).

1.2 IFFT and FFT utilization: A/D realization of OFDM

A significant breakthrough in OFDM applicability was presented by Weinstein and Ebert in 1971 (Weinstein & Ebert, 1971). First, DFT and inverse DFT (IDFT) techniques were applied

to OFDM implementation to perform base-band parallel sub-channel modulations and demodulations (or multiplexing and demultiplexing) (Weinstein & Ebert, 1971). This study provided an effective discrete-time signal processing method to simultaneously modulate (and demodulate) signals transmitted (and received) on various sub-channels without requiring the implementation of a bank of sub-carrier modulators with many analog multipliers and oscillators. Meanwhile, ISI can be significantly reduced by inserting a guard time-interval (GI) in between any two consecutive OFDM symbols and by applying a raised-cosine windowing method to the time-domain (TD) signals (Weinstein & Ebert, 1971). Although the system studied in this work cannot always maintain orthogonality among sub-carriers when operated over a time-dispersive channel, the application of IDFT and DFT to OFDM communication is not only a crucial contribution but also a critical driving force for commercial applicability of recent wireless OFDM communication because the fast algorithms of IDFT and DFT, i.e., inverse fast Fourier transform (IFFT) and fast Fourier transform (FFT), have been commercialized and popularly implemented with ASICs or sub-functions on DSPs.

1.3 Cyclic prefix

Orthogonality among sub-carriers cannot be maintained when an OFDM system operates over a time-dispersive channel. This problem was first addressed by Peled and Ruiz in 1980 (Peled & Ruiz, 1980). Rather than inserting a blank GI between any two consecutive OFDM symbols, which was the method employed in the previous study (Weinstein & Ebert, 1971), a cyclic extension of an OFDM block is inserted into the original GI as a prefix to an information-bearing OFDM block. The adopted cyclic prefix (CP) effectively converts the linear convolution of the transmitted symbol and the channel impulse response (CIR) into the cyclic convolution; thus, orthogonality among sub-carriers can be maintained with channel time-dispersion if the CP is sufficiently longer than the CIR. However, energy efficiency is inevitably sacrificed, as the CPs convey no desired information.

1.4 Applications

OFDM technology is currently employed in the European digital audio broadcasting (DAB) standard (DAB, 1995). In addition, digital TV broadcasting applications based on OFDM technology have been under comprehensive investigation (DVB, 1996; Couasnon et al., 1994; Marti et al., 1993; Moeneclaey & Bladel, 1993; Tourtier et al., 1993). Furthermore, OFDM technology in conjunction with other multiple-access techniques, in particular code-division multiple-access (CDMA) techniques, for mobile communications has also been the focus of a variety of research efforts (Hara & Prasad, 1997; Sourour & Nakagawa, 1996; Kondo & Milstein, 1996; Reiners & Rohling, 1994; Fazel, 1994). For those employed in wireline environments, OFDM communication systems are often called "Discrete Multi-Tone" (DMT) communications, which have also attracted a great deal of research attention as a technology that effectively achieves high-rate transmission on currently existing telephone networks (Bingham, 1990; Young et al., 1996; Chow, 1993; Tu, 1991). One of the major advantages of the OFDM technique is its robustness with multipath reception. OFDM applications often are expected to operate in a severely frequency-selective environment. Therefore, OFDM communication has recently been selected for various broadband mobile communications, e.g., 4G mobile communications. This chapter will focus on such applications.

1.5 System description and signal modelling

The primary idea behind OFDM communication is dividing an occupied frequency band into many parallel sub-channels to deliver information simultaneously. By maintaining sufficiently narrow sub-channel bandwidths, the signal propagating through an individual sub-channel experiences roughly frequency-flat (i.e., frequency-nonselective) channel fades. This arrangement can significantly reduce the complexity of the subsequent equalization sub-system. In particular, current broadband wireless communications are expected to be able to operate in severe multipath fading environments in which long delay spreads inherently exist because the signature/chip duration has become increasingly shorter. To enhance spectral (or bandwidth) efficiency, the spectra of adjacent sub-channels are set to overlap with one another. Meanwhile, the orthogonality among sub-carriers is maintained by setting the sub-carrier spacing (i.e., the frequency separation between two consecutive sub-carriers) to the reciprocal of an OFDM block duration.

By taking advantage of a CP, the orthogonality can be prevented from experiencing ICI even for transmission over a multipath channel (Peled & Ruiz, 1980). Although several variants of OFDM communication systems exist (Bingham, 1990; Weinstein & Ebert, 1971; Floch et al., 1995), CP-OFDM (Peled & Ruiz, 1980) is primarily considered in this section due to its popularity. A CP is obtained from the tail portion of an OFDM block and then prefixed into a transmitted block, as shown in Fig. 1.

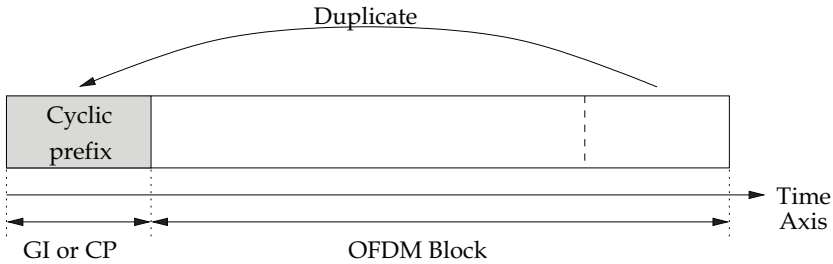


Fig. 1. An OFDM symbol consisting of a CP and an information-bearing OFDM block.

A portion of the transmitted OFDM symbol becomes periodic. The CP insertion converts the linear convolution of the CIR and the transmitted symbol into the circular convolution of the two. Therefore, CPs can avoid both ISI and ICI (Bingham, 1990). In this fundamental section, the following assumptions are made for simplicity: (1) a cyclic prefix is used; (2) the CIR length does not exceed the CP length; (3) the received signal can be perfectly synchronized; (4) noise is complex-valued, additive, white Gaussian noise (AWGN); and (5) channel time-variation is slow, so the channel can be considered to be constant or static within a few OFDM symbols.

1.5.1 Continuous-time model

A continuous-time base-band equivalent representation of an OFDM transceiver is depicted in Fig. 2. The OFDM communication system under study consists of N sub-carriers that occupy a total bandwidth of $B = \frac{1}{T_s}$ Hz. The length of an OFDM symbol is set to T_{sym} seconds; moreover, an OFDM symbol is composed of an OFDM block of length $T = NT_s$ and a CP of length T_g . The transmitting filter on the k th sub-carrier can be written as

$$p_k(t) = \begin{cases} \frac{1}{\sqrt{T}} e^{j2\pi \frac{B}{N} k(t-T_g)} & 0 \leq t \leq T_{sym} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $T_{sym} = T + T_g$. Note that $p_k(t) = p_k(t+T)$ when t is within the guard interval $[0, T_g]$. It can be seen from Equation 1 that $p_k(t)$ is a rectangular pulse modulated by a sub-carrier with frequency $k \cdot \frac{B}{N}$. The transmitted signal $s_i(t)$ for the i th OFDM symbol can thus be obtained by summing over all modulated signals, i.e.,

$$s_i(t) = \sum_{k=0}^{N-1} X_{k,i} p_k(t - iT_{sym}), \quad (2)$$

where $X_{0,i}, X_{1,i}, \dots, X_{N-1,i}$ are complex-valued information-bearing symbols, whose values are often mapped according to quaternary phase-shift keying (QPSK) or quadrature amplitude modulation (QAM). Therefore, the transmitted signal $s(t)$ can be considered to be a sequence of OFDM symbols, i.e.,

$$\begin{aligned} s(t) &= \sum_{i=-\infty}^{\infty} s_i(t) \\ &= \sum_{i=-\infty}^{\infty} \sum_{k=0}^{N-1} X_{k,i} p_k(t - iT_{sym}). \end{aligned} \quad (3)$$

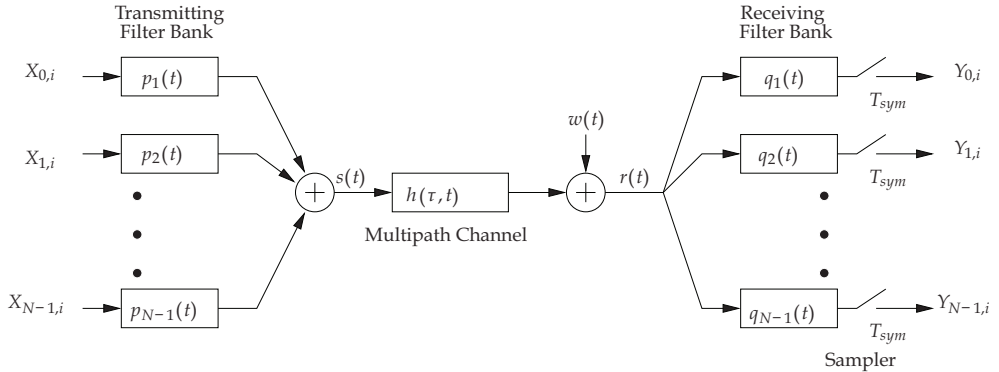


Fig. 2. Continuous-time base-band equivalent representation of an OFDM transceiver.

If the length of the CIR $h(\tau, t)$ does not exceed the CP length T_g , the received signal $r(t)$ can be written as

$$\begin{aligned} r(t) &= (h * s)(t) + w(t) \\ &= \int_0^{T_g} h(\tau, t) s(t - \tau) d\tau + w(t), \end{aligned} \quad (4)$$

where the operator “*” represents the linear convolution and $w(t)$ is an AWGN.

At the receiving end, a bank of filters is employed to match the last part $[T_g, T_{sym}]$ of the transmitted waveforms $p_k(t)$ on a subchannel-by-subchannel basis. By taking advantage of

matched filter (MF) theory, the receiving filter on the k th sub-channel can be designed to have the following impulse response:

$$q_k(t) = \begin{cases} p_k^*(T_{sym} - t), & 0 \leq t < T = T_{sym} - T_g \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Because the CP can effectively separate symbol dispersion from preceding or succeeding symbols, the sampled outputs of the receiving filter bank convey negligible ISI. The time index i can be dropped for simplicity because the following derivations address the received signals on a symbol-by-symbol basis and the ISI is considered to be negligible. Using Equations 3, 4 and 5, the sampled output of the k th receiving MF can be written as

$$\begin{aligned} Y_k &= (r * q_k)(t) \Big|_{t=T_{sym}} \\ &= \int_{-\infty}^{\infty} r(\zeta) q_k(T_{sym} - \zeta) d\zeta \\ &= \int_{T_g}^{T_{sym}} \left(\int_0^{T_g} h(\tau, t) s(\zeta - \tau) d\tau + w(\zeta) \right) p_k^*(\zeta) d\zeta \\ &= \int_{T_g}^{T_{sym}} \left(\int_0^{T_g} h(\tau, t) \left[\sum_{l=0}^{N-1} X_l p_l(\zeta - \tau) \right] d\tau \right) p_k^*(\zeta) d\zeta + \int_{T_g}^{T_{sym}} w(\zeta) p_k^*(\zeta) d\zeta. \end{aligned} \quad (6)$$

It is assumed that although the CIR is time-varying, it does not significantly change within a few OFDM symbols. Therefore, the CIR can be further represented as $h(\tau)$. Equation 6 can thus be rewritten as

$$Y_k = \sum_{l=0}^{N-1} X_l \int_{T_g}^{T_{sym}} \left(\int_0^{T_g} h(\tau) p_l(\zeta - \tau) d\tau \right) p_k^*(\zeta) d\zeta + \int_{T_g}^{T_{sym}} w(\zeta) p_k^*(\zeta) d\zeta. \quad (7)$$

From Equation 7, if $T_g < \zeta < T_{sym}$ and $0 < \tau < T_g$, then $0 < \zeta - \tau < T_{sym}$. Therefore, by substituting Equation 1 into Equation 7, the inner-most integral of Equation 7 can be reformulated as

$$\begin{aligned} \int_0^{T_g} h(\tau) p_l(\zeta - \tau) d\tau &= \int_0^{T_g} h(\tau) \frac{e^{j2\pi l(\zeta - \tau - T_g)B/N}}{\sqrt{T}} d\tau \\ &= \frac{e^{j2\pi l(\zeta - T_g)B/N}}{\sqrt{T}} \int_0^{T_g} h(\tau) e^{-j2\pi l\tau B/N} d\tau, \quad T_g < \zeta < T_{sym}. \end{aligned} \quad (8)$$

Furthermore, the integration in Equation 8 can be considered to be the channel weight of the l th sub-channel, whose sub-carrier frequency is $f = lB/N$, i.e.,

$$H_l = H \left(l \frac{B}{N} \right) = \int_0^{T_g} h(\tau) e^{-j2\pi l\tau B/N} d\tau, \quad (9)$$

where $H(f)$ denotes the channel transfer function (CTF) and is thus the Fourier transform of $h(\tau)$. The output of the k th receiving MF can therefore be rewritten as

$$\begin{aligned} Y_k &= \sum_{l=0}^{N-1} X_l \int_{T_g}^{T_{sym}} \frac{e^{j2\pi l(\zeta-T_g)B/N}}{\sqrt{T}} H_l p_k^*(\zeta) d\zeta + \int_{T_g}^{T_{sym}} w(\zeta) p_k^*(\zeta) d\zeta \\ &= \sum_{l=0}^{N-1} X_l H_l \int_{T_g}^{T_{sym}} p_l(\zeta) p_k^*(\zeta) d\zeta + W_k, \end{aligned} \quad (10)$$

where

$$W_k = \int_{T_g}^{T_{sym}} w(\zeta) p_k^*(\zeta) d\zeta.$$

The transmitting filters $p_k(t)$, $k = 0, 1, \dots, N-1$ employed here are mutually orthogonal, i.e.,

$$\begin{aligned} \int_{T_g}^{T_{sym}} p_l(t) p_k^*(t) dt &= \int_{T_g}^{T_{sym}} \frac{e^{j2\pi l(t-T_g)B/N}}{\sqrt{T}} \frac{e^{-j2\pi k(t-T_g)B/N}}{\sqrt{T}} dt \\ &= \delta[k-l], \end{aligned} \quad (11)$$

where

$$\delta[k-l] = \begin{cases} 1 & k=l \\ 0 & \text{otherwise} \end{cases}$$

is the Kronecker delta function. Therefore, Equation 10 can be reformulated as

$$Y_k = H_k X_k + W_k, \quad k = 0, 1, \dots, N-1, \quad (12)$$

where W_k is the AWGN of the k th sub-channel. As a result, the OFDM communication system can be considered to be a set of parallel frequency-flat (frequency-nonselective) fading sub-channels with uncorrelated noise, as depicted in Fig. 3.

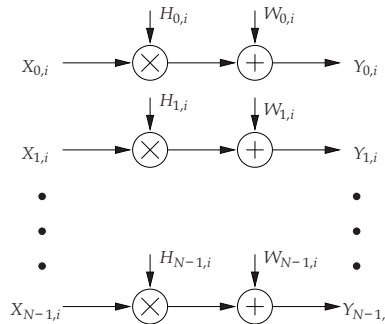


Fig. 3. OFDM communication is converted to transmission over parallel frequency-flat sub-channels.

1.5.2 Discrete-time model

A fully discrete-time representation of the OFDM communication system studied here is depicted in Fig. 4. The modulation and demodulation operations in the continuous-time model have been replaced by IDFT and DFT operations, respectively, and the channel has been replaced by a discrete-time channel.

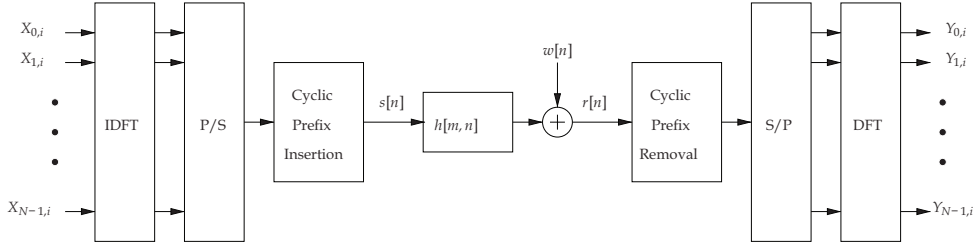


Fig. 4. Discrete-time representation of a base-band equivalent OFDM communication system.

If the CP is longer than the CIR, then the linear convolution operation can be converted to a cyclic convolution. The cyclic convolution is denoted as ' \otimes ' in this chapter. The i th block of the received signals can be written as

$$\begin{aligned} \mathbf{Y}_i &= \text{DFT}_N \left\{ \text{IDFT}_N \{ \mathbf{X}_i \} \otimes \mathbf{h}_i + \mathbf{w}_i \right\} \\ &= \text{DFT}_N \left\{ \text{IDFT}_N \{ \mathbf{X}_i \} \otimes \mathbf{h}_i \right\} + \mathbf{W}_i, \end{aligned} \quad (13)$$

where $\mathbf{Y}_i = [Y_{0,i} \ Y_{1,i} \ \dots \ Y_{N-1,i}]^T$ is an $N \times 1$ vector, and its elements represent N demodulated symbols; $\mathbf{X}_i = [X_{0,i} \ X_{1,i} \ \dots \ X_{N-1,i}]^T$ is an $N \times 1$ vector, and its elements represent N transmitted information-bearing symbols; $\mathbf{h}_i = [h_{0,i} \ h_{1,i} \ \dots \ h_{N-1,i}]^T$ is an $N \times 1$ vector, and its elements represent the CIR padded with sufficient zeros to have N dimensions; and $\mathbf{w}_i = [w_{0,i} \ w_{1,i} \ \dots \ w_{N-1,i}]^T$ is an $N \times 1$ vector representing noise. Because the noise is assumed to be white, Gaussian and circularly symmetric, the noise term

$$\mathbf{W}_i = \text{DFT}_N(\mathbf{w}_i) \quad (14)$$

represents uncorrelated Gaussian noise, and $W_{k,i}$ and $w_{n,i}$ can be proven to have the same variance according to the Central Limit Theorem (CLT). Furthermore, if a new operator " \odot " is defined to be element-by-element multiplication, Equation 13 can be rewritten as

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i \odot \text{DFT}_N \{ \mathbf{h}_i \} + \mathbf{W}_i \\ &= \mathbf{X}_i \odot \mathbf{H}_i + \mathbf{W}_i, \end{aligned} \quad (15)$$

where $\mathbf{H}_i = \text{DFT}_N \{ \mathbf{h}_i \}$ is the CTF. As a result, the same set of parallel frequency-flat subchannels with noise as presented in the continuous-time model can be obtained.

Both the aforementioned continuous-time and discrete-time representations provide insight and serve the purpose of providing a friendly first step or entrance point for beginning readers. In my personal opinion, researchers that have more experience in communication fields may be more comfortable with the continuous-time model because summations, integrations and convolutions are employed in the modulation, demodulation and (CIR)

filtering processes. Meanwhile, researchers that have more experience in signal processing fields may be more comfortable with the discrete-time model because vector and matrix operations are employed in the modulation, demodulation and (CIR) filtering processes. Although the discrete-time model may look neat, clear and reader-friendly, several presumptions should be noted and kept in mind. It is assumed that the symbol shaping is rectangular and that the frequency offset, ISI and ICI are negligible. The primary goal of this chapter is to highlight concepts and provide insight to beginning researchers and practical engineers rather than covering theories or theorems. As a result, the derivations shown in Sections 3 and 4 are close to the continuous-time representation, and those in Sections 5 and 6 are derived from the discrete-time representation.

2. Introduction to channel estimation on wireless OFDM communications

2.1 Preliminary

In practice, effective channel estimation (CE) techniques for coherent OFDM communications are highly desired for demodulating or detecting received signals, improving system performance and tracking time-varying multipath channels, especially for mobile OFDM because these techniques often operate in environments where signal reception is inevitably accompanied by wide Doppler spreads caused by dynamic surroundings and long multipath delay spreads caused by time-dispersion. Significant research efforts have focused on addressing various CE and subsequent equalization problems by estimating sub-channel gains or the CIR. CE techniques in OFDM systems often exploit several pilot symbols transceived at given locations on the frequency-time grid to determine the relevant channel parameters. Several previous studies have investigated the performance of CE techniques assisted by various allocation patterns of the pilot/training symbols (Coleri et al., 2002; Li et al., 2002; Yeh & Lin, 1999; Negi & Cioffi, 1998). Meanwhile, several prior CEs have simultaneously exploited both time-directional and frequency-directional correlations in the channel under investigation (Hoeher et al., 1997; Wilson et al., 1994; Hoeher, 1991). In practice, these two-dimensional (2D) estimators require 2D Wiener filters and are often too complicated to be implemented. Moreover, it is difficult to achieve any improvements by using a 2D estimator, while significant computational complexity is added (Sandell & Edfors, 1996). As a result, serially exploiting the correlation properties in the time and frequency directions may be preferred (Hoeher, 1991) for reduced complexity and good estimation performance. In mobile environments, channel tap-weighting coefficients often change rapidly. Thus, the comb-type pilot pattern, in which pilot symbols are inserted and continuously transmitted over specific pilot sub-channels in all OFDM blocks, is naturally preferred and highly desirable for effectively and accurately tracking channel time-variations (Negi & Cioffi, 1998; Wilson et al., 1994; Hoeher, 1991; Hsieh & Wei, 1998).

Several methods for allocating pilots on the time-frequency grid have been studied (Tufvesson & Maseng, 1997). Two primary pilot assignments are depicted in Fig. 5: the block-type pilot arrangement (BTPA), shown in Fig. 5(a), and the comb-type pilot arrangement (CTPA), shown in Fig. 5(b). In the BTPA, pilot signals are assigned in specific OFDM blocks to occupy all sub-channels and are transmitted periodically. Both in general and in theory, BTPA is more suitable in a slowly time-varying, but severely frequency-selective fading environment. No interpolation method in the FD is required because the pilot block occupies the whole band. As a result, the BTPA is relatively insensitive to severe

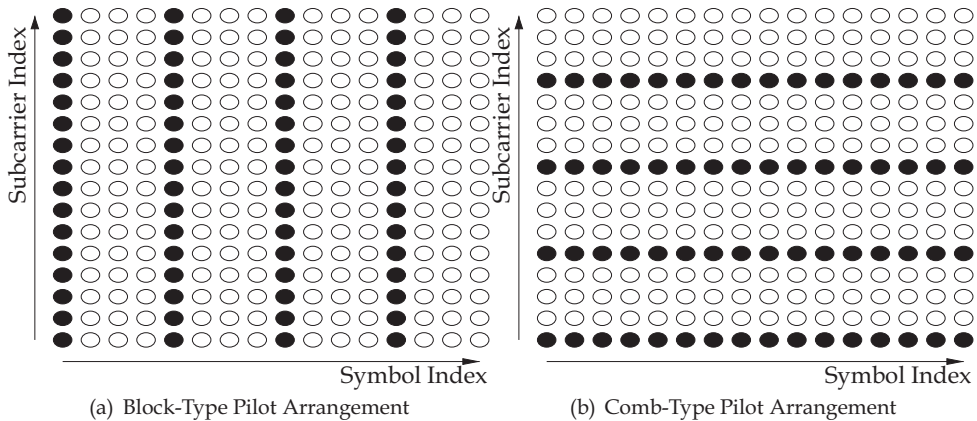


Fig. 5. Two primary pilot assignment methods

frequency selectivity in a multipath fading channel. Estimates of the CIR can usually be obtained by least-squares (LS) or minimum-mean-square-error (MMSE) estimations conducted with assistance from the pilot symbols (Edfors et al., 1996; Van de Beek et al., 1995).

In the CTPA, pilot symbols are often uniformly distributed over all sub-channels in each OFDM symbol. Therefore, the CTPA can provide better resistance to channel time-variations. Channel weights on non-pilot (data) sub-channels have to be estimated by interpolating or smoothing the estimates of the channel weights obtained on the pilot sub-channels (Zhao & Huang, 1997; Rinne & Renfors, 1996). Therefore, the CTPA is, both in general and in theory, sensitive to the frequency-selectivity of a multipath fading channel. The CTPA is adopted to assist the CE conducted in each OFDM block in Sections 3 and 4, while the BTPA is discussed in Section 5.

2.2 CTPA-based CE

Conventional CEs assisted by comb-type pilot sub-channels are often performed completely in the frequency domain (FD) and include two steps: jointly estimating the channel gains on all pilot sub-channels and smoothing the obtained estimates to interpolate the channel gains on data (non-pilot) sub-channels. The CTPA CE technique (Hsieh & Wei, 1998) and the pilot-symbol-assisted modulation (PSAM) CE technique (Edfors et al., 1998) have been shown to be practical and applicable methods for mobile OFDM communication because their ability to track rapidly time-varying channels is much better than that of a BTPA CE technique. Several modified variants for further improvements and for complexity or rank reduction by means of singular-value-decomposition (SVD) techniques have been investigated previously (Hsieh & Wei, 1998; Edfors et al., 1998; Seller, 2004; Edfors et al., 1996; Van de Beek et al., 1995; Park et al., 2004). In addition, a more recent study has proposed improving CE performance by taking advantage of presumed slowly varying properties in the delay subspace (Simeone et al., 2004). This technique employs an intermediate step between the LS pilot sub-channel estimation step and the data sub-channel interpolation step in conventional CE approaches (Hsieh & Wei, 1998; Edfors et al., 1998; Seller, 2004; Edfors et al., 1996; Van de Beek et al., 1995; Park et al., 2004) to track the delay subspace to improve the accuracy of the pilot sub-channel estimation. However, this

technique is based on the strong assumption that the multipath delays are slowly time-varying and can easily be estimated separately from the channel gain estimation. A prior channel estimation study (Minn & Bhargava, 2000) also exploited CTPA and TD CE. The proposed technique (Minn & Bhargava, 2000) was called the Frequency-Pilot-Time-Average (FPTA) method. However, time-averaging over a period that may be longer than the coherence time of wireless channels to suppress interference not only cannot work for wireless applications with real-time requirements but may also be impractical in a mobile channel with a short coherence time. A very successful technique that takes advantage of TD CE has been proposed (Minn & Bhargava, 1999). However, this technique focused on parameter estimation to transmit diversity using space-time coding in OFDM systems, and the parameter settings were not obtained from any recent mobile communication standards. To make fair comparisons of the CE performance and to avoid various diversity or space-time coding methods, only uncoded OFDM with no diversity is addressed in this chapter.

The CTPA is also employed as the framework of the technique studied in Sections 3 and 4 because of its effectiveness in mobile OFDM communications with rapidly time-varying, frequency-selective fading channels. A least-squares estimation (LSE) approach is performed serially on a block-by-block basis in the TD, not only to accurately estimate the CIR but also to effectively track rapid CIR variations. In fact, a generic estimator is thus executed on each OFDM block without assistance from a priori channel information (e.g., correlation functions in the frequency and/or in the time directions) and without increasing computational complexity.

Many previous studies (Edfors et al., 1998; Seller, 2004; Edfors et al., 1996; Van de Beek et al., 1995; Simeone et al., 2004) based on CTPA were derived under the assumption of perfect timing synchronization. In practice, some residual timing error within several sampling durations inevitably occurs during DFT demodulation, and this timing error leads to extra phase errors that phase-rotate demodulated symbols. Although a method that solves this problem in conventional CTPA OFDM CEs has been studied (Hsieh & Wei, 1998; Park et al., 2004), this method can work only under some special conditions (Hsieh & Wei, 1998). Compared with previous studies (Edfors et al., 1998; Seller, 2004; Edfors et al., 1996; Van de Beek et al., 1995; Simeone et al., 2004), the studied technique can be shown to achieve better resistance to residual timing errors because it does not employ a priori channel information and thus avoids the model mismatch and extra phase rotation problems that result from residual timing errors. Also, because the studied technique performs ideal data sub-channel interpolation with a domain-transformation approach, it can effectively track extra phase rotations with no phase lag.

2.3 BTPA-based CE

Single-carrier frequency-division multiple-access (SC-FDMA) communication was selected for the long-term evolution (LTE) specification in the third-generation partnership project (3GPP). SC-FDMA has been the focus of research and development because of its ability to maintain a low peak-to-average power ratio (PAPR), particularly in the uplink transmission, which is one of a few problems in recent 4G mobile communication standardization. Meanwhile, SC-FDMA can maintain high throughput and low equalization complexity like orthogonal frequency-division multiple access (OFDMA) (Myung et al., 2006). Moreover, SC-FDMA can be thought of as an OFDMA with DFT pre-coded or pre-spread inputs. In a SC-FDMA uplink scenario, information-bearing symbols in the TD from any individual user terminal are pre-coded (or pre-spread) with a DFT. The DFT-spread resultant symbols can

be transformed into the FD. Finally, the DFT-spread symbols are fed into an IDFT multiplexer to accomplish FDM.

Although the CTPA is commonly adopted in wireless communication applications, such as IEEE 802.11a, IEEE 802.11g, IEEE 802.16e and the EU-IST-4MORE project, the BTPA is employed in the LTE. As shown in the LTE specification, 7 symbols form a slot, and 20 slots form a frame that spans 10 ms in the LTE uplink transmission. In each slot, the 4th symbol is used to transmit a pilot symbol. Section 5 employs BTPA as the framework to completely follow the LTE specifications. A modified Kalman filter- (MKF-) based TD CE approach with fast fading channels has been proposed previously (Han et al., 2004). The MKF-based TD CE tracks channel variations by taking advantage of MKF and TD MMSE equalizers. A CE technique that also employs a Kalman filter has been proposed (Li et al., 2008). Both methods successfully address the CE with high Doppler spreads.

The demodulation reference signal adopted for CE in LTE uplink communication is generated from Zadoff-Chu (ZC) sequences. ZC sequences, which are generalized chirp-like poly-phase sequences, have some beneficial properties according to previous studies (Ng et al., 1998; Popovic, 1992). ZC sequences are also commonly used in radar applications and as synchronization signals in LTE, e.g., random access and cell search (Levanon & Mozeson, 2004; LTE, 2009). A BTPA-based CE technique is discussed in great detail in Section 5.

2.4 TD-redundancy-based CE

Although the mobile communication applications mentioned above are all based on cyclic-prefix OFDM (CP-OFDM) modulation techniques, several encouraging contributions have investigated some alternatives, e.g., zero-padded OFDM (ZP-OFDM) (Muquest et al., 2002; Muquet et al., 2000) and pseudo-random-postfix OFDM (PRP-OFDM) (Muck et al., 2006; 2005; 2003) to replace the TD redundancy with null samples or known/pre-determined sequences. It has been found that significant improvements over CP-OFDM can be realized with either ZP-OFDM or PRP-OFDM (Muquest et al., 2002; Muquet et al., 2000; Muck et al., 2006; 2005; 2003). In previous works, ZP-OFDM has been shown to maintain symbol recovery irrespective of null locations on a multipath channel (Muquest et al., 2002; Muquet et al., 2000). Meanwhile, PRP-OFDM replaces the null samples originally inserted between any two OFDM blocks in ZP-OFDM by a known sequence. Thus, the receiver can use the a priori knowledge of a fraction of transmitted blocks to accurately estimate the CIR and effectively reduce the loss of transmission rate with frequent, periodic training sequences (Muck et al., 2006; 2005; 2003). A more recent OFDM variant, called Time-Domain Synchronous OFDM (TDS-OFDM) was investigated in terrestrial broadcasting applications (Gui et al., 2009; Yang et al., 2008; Zheng & Sun, 2008; Liu & Zhang, 2007; Song et al., 2005). TDS-OFDM works similarly to the PRP-OFDM and also belongs to this category of CEs assisted by TD redundancy.

Several research efforts that address various PRP-OFDM CE and/or subsequent equalization problems have been undertaken (Muck et al., 2006; 2005; 2003; Ma et al., 2006). However, these studies were performed only in the context of a wireless local area network (WLAN), in which multipath fading and Doppler effects are not as severe as in mobile communication. In addition, the techniques studied in previous works (Muck et al., 2006; 2005; 2003; Ma et al., 2006) take advantage of a time-averaging method to replace statistical expectation operations and to suppress various kinds of interference, including inter-block interference (IBI) and ISI. However, these moving-average-based interference suppression methods investigated in the previous studies (Muck et al., 2006; 2005; 2003; Ma et al., 2006)

cannot function in the mobile environment because of rapid channel variation and real-time requirements. In fact, it is difficult to design an effective moving-average filter (or an integrate-and-dump (I/D) filter) for the previous studies (Muck et al., 2006; 2005; 2003; Ma et al., 2006) because the moving-average filter must have a sufficiently short time-averaging duration (i.e., sufficiently short I/D filter impulse response) to accommodate both the time-variant behaviors of channel tap-weighting coefficients and to keep the a priori statistics of the PRP unchanged for effective CE and must also have a sufficiently long time-averaging duration (i.e., sufficiently long I/D filter impulse response) to effectively suppress various kinds of interference and reduce AWGN.

A previous work (Ohno & Giannakis, 2002) investigated an optimum training pattern for generic block transmission over time-frequency selective channels. It has been proven that the TD training sequences must be placed with equal spacing to minimize mean-square errors. However, the work (Ohno & Giannakis, 2002) was still in the context of WLAN and broadcasting applications, and no symbol recovery method was studied. As shown in Section 6, the self-interference that occurs with symbol recovery and signal detection must be further eliminated by means of the SIC method.

3. Frequency-domain channel estimation based on comb-type pilot arrangement

3.1 System description

The block diagram of the OFDM transceiver under study is depicted in Fig. 6. Information-bearing bits are grouped and mapped according to Gray encoding to become multi-amplitude-multi-phase symbols. After pilot symbol insertion, the block of data $\{X_k, k = 0, 1, \dots, N-1\}$ is then fed into the IDFT (or IFFT) modulator. Thus, the modulated symbols $\{x_n, n = 0, 1, \dots, N-1\}$ can be expressed as

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}, \quad n = 0, 1, \dots, N-1, \quad (16)$$

where N is the number of sub-channels. In the above equation, it is assumed that there are no virtual sub-carriers, which provide guard bands, in the studied OFDM system. A CP is arranged in front of an OFDM symbol to avoid ISI and ICI, and the resultant symbol $\{x_{cp,n}, n = -L, -L+1, \dots, N-1\}$ can thus be expressed as

$$x_{cp,n} = \begin{cases} x_{N+n} & n = -L, -L+1, \dots, -1 \\ x_n & n = 0, 1, \dots, N-1, \end{cases} \quad (17)$$

where L denotes the number of CP samples. The transmitted signal is then fed into a multipath fading channel with CIR $h[m, n]$. The received signal can thus be represented as

$$y_{cp}[n] = x_{cp}[n] \otimes h[m, n] + w[n], \quad (18)$$

where $w[n]$ denotes the AWGN. The CIR $h[m, n]$ can be expressed as (Steele, 1999)

$$h[m, n] = \sum_{i=0}^{M-1} \alpha_i e^{j2\pi v_i n T_s} \delta[mT_s - \tau_i], \quad (19)$$

where M denotes the number of resolvable propagation paths, α_i represents the i th complex channel weight of the CIR, ν_i denotes the maximum Doppler frequency on the i th resolvable propagation path, m is the index in the delay domain, n is the time index, and τ_i denotes the delay of the i th resolvable path.

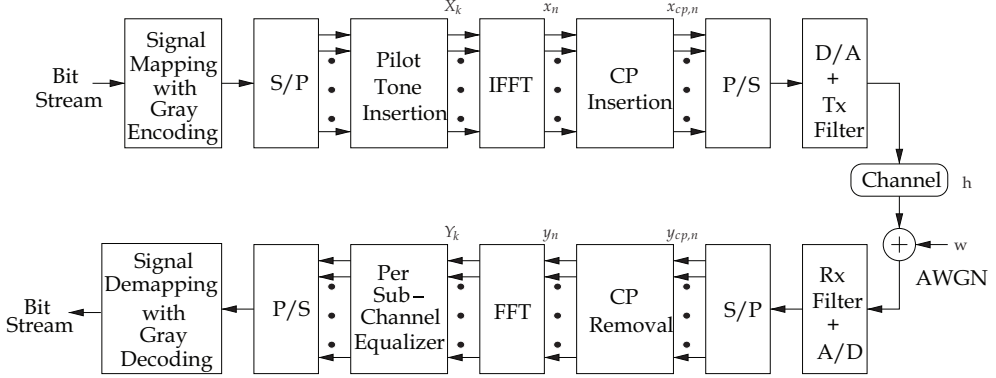


Fig. 6. A base-band equivalent block diagram of the studied OFDM transceiver.

After the CP portion is effectively removed from $y_{cp,n}$, the received samples y_n are sifted and fed into the DFT demodulator to simultaneously demodulate the signals propagating through the multiple sub-channels. The demodulated symbol obtained on the k th sub-channel can thus be written as

$$Y_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y_n e^{-j2\pi kn/N}, \quad k = 0, 1, \dots, N-1. \quad (20)$$

If the CP is sufficiently longer than the CIR, then the ISI among OFDM symbols can be neglected. Therefore, Y_k can be reformulated as (Zhao & Huang, 1997; Hsieh & Wei, 1998)

$$Y_k = X_k H_k + I_k + W_k, \quad k = 0, 1, \dots, N-1, \quad (21)$$

where

$$\begin{aligned} H_k &= \sqrt{N} \sum_{i=0}^{M-1} \alpha_i e^{j\nu_i T} \frac{\sin(\pi \nu_i T)}{\pi \nu_i T} e^{-j\frac{2\pi \tau_i}{N} k}, \\ I_k &= \frac{1}{\sqrt{N}} \sum_{i=0}^{M-1} \alpha_i \sum_{\substack{k'=0 \\ k' \neq k}}^{N-1} X(k') \frac{1 - e^{j2\pi(\nu_i T + k' - k)}}{1 - e^{j\frac{2\pi}{N}(\nu_i T + k' - k)}} e^{-j\frac{2\pi \tau_i}{N} k'}, \quad k = 0, 1, \dots, N-1 \end{aligned} \quad (22)$$

and $\{W_k, k = 0, 1, \dots, N-1\}$ is the Fourier transform of $\{w_n, n = 0, 1, \dots, N-1\}$.

The symbols $\{Y_{p,k}\}$ received on the pilot sub-channels can be obtained from $\{Y_k, k = 0, 1, \dots, N-1\}$, the channel weights on the pilot sub-channels $\{H_{p,k}\}$ can be estimated, and then the channel weights on the data (non-pilot) sub-channels can be obtained by interpolating or smoothing the obtained estimates of the pilot sub-channel weights $H_{p,k}$. The transmitted information-bearing symbols $\{X_k, k = 0, 1, \dots, N-1\}$ can be recovered by simply dividing the received symbols by the corresponding channel weights, i.e.,

$$\hat{X}_k = \frac{Y_k}{\hat{H}_k}, \quad k = 0, 1, \dots, N-1, \quad (22)$$

where \hat{H}_k is an estimate of H_k . Eventually, the source binary data may be reconstructed by means of signal demapping.

3.2 Pilot sub-channel estimation

In the CTPA, the N_p pilot signals $X_{p,m}$, $m = 0, 1, \dots, N_p - 1$ are inserted into the FD transmitted symbols X_k , $k = 0, 1, \dots, N-1$ with equal separation. In other words, the total N sub-carriers are divided into N_p groups, each of which contains $Q = N/N_p$ contiguous sub-carriers. Within any group of sub-carriers, the first sub-carrier, with the lowest central frequency, is adopted to transmit pilot signals. The value of $\rho = Q^{-1}$ denotes the pilot density employed in the OFDM communication studied here. The pilot density ρ represents the portion of the entire bandwidth that is employed to transmit the pilots, and it must be as low as possible to maintain sufficiently high bandwidth efficiency. However, the Nyquist sampling criterion sets a lower bound on the pilot density ρ that allows the CTF to be effectively reconstructed with a subcarrier-domain (i.e., FD) interpolation approach. The OFDM symbol transmitted over the k th sub-channel can thus be expressed as

$$X_k = \begin{cases} X_{mQ+l} \\ X_{p,m'} \\ \text{information} \end{cases} \quad \begin{cases} l = 0, \\ l = 1, 2, \dots, Q-1. \end{cases} \quad (23)$$

The pilot signals $\{X_{p,m}, m = 0, 1, \dots, N_p - 1\}$ can either be a common complex value or sifted from a pseudo-random sequence.

The channel weights on the pilot sub-channels can be written in vector form, i.e.,

$$\mathbf{H}_p = \begin{bmatrix} H_p(0) & H_p(1) & \dots & H_p(N_p - 1) \end{bmatrix}^T \\ = \begin{bmatrix} H(0) & H(Q) & \dots & H((N_p - 1)Q) \end{bmatrix}^T. \quad (24)$$

The received symbols on the pilot sub-channels obtained after the FFT demodulation can be expressed as

$$\mathbf{Y}_p = \begin{bmatrix} Y_{p,0} & Y_{p,1} & \dots & Y_{p,N_p-1} \end{bmatrix}^T. \quad (25)$$

Moreover, \mathbf{Y}_p can be rewritten as

$$\mathbf{Y}_p = \mathbf{X}_p \cdot \mathbf{H}_p + \mathbf{I}_p + \mathbf{W}_p, \quad (26)$$

where

$$\mathbf{X}_p = \begin{bmatrix} X_p(0) & & & \mathbf{0} \\ & \ddots & & \\ & & & X_p(N_p - 1) \\ \mathbf{0} & & & \end{bmatrix},$$

\mathbf{I}_p denotes the ICI vector and \mathbf{W}_p denotes the AWGN of the pilot sub-channels.

In conventional CTPA-based CE methods, the estimates of the channel weights of the pilot sub-channels can be obtained by means of the LS CE, i.e.,

$$\begin{aligned}\hat{\mathbf{H}}_{LS} &= \left[H_{p,LS}(0) \ H_{p,LS}(1) \ \cdots \ H_{p,LS}(N_p-1) \right]^T \\ &= \left(\mathbf{X}_p^H \mathbf{X}_p \right)^{-1} \mathbf{X}_p^H \mathbf{Y}_p = \mathbf{X}_p^{-1} \mathbf{Y}_p \\ &= \left[\frac{Y_p(0)}{X_p(0)} \ \frac{Y_p(1)}{X_p(1)} \ \cdots \ \frac{Y_p(N_p-1)}{X_p(N_p-1)} \right]^T.\end{aligned}\quad (27)$$

Although the aforementioned LS CE $\hat{\mathbf{H}}_{LS}$ enjoys low computational complexity, it suffers from noise enhancement problems, like the zero-forcing equalizer discussed in textbooks.

The MMSE criterion is adopted in CE and equalization techniques, and it exhibits better CE performance than the LS CE in OFDM communications assisted by block pilots (Van de Beek et al., 1995). The main drawback of the MMSE CE is its high complexity, which grows exponentially with the size of the observation samples. In a previous study (Edfors et al., 1996), a low-rank approximation was applied to a linear minimum-mean-square-error (LMMSE) CE assisted by FD correlation. The key idea to reduce the complexity is using the singular-value-decomposition (SVD) technique to derive an optimal low-rank estimation, the performance of which remains essentially unchanged. The MMSE CE performed on the pilot sub-channels is formulated as follows (Edfors et al., 1996):

$$\begin{aligned}\hat{\mathbf{H}}_{LMMSE} &= \mathbf{R}_{\hat{H}_{LS}\hat{H}_{LS}} \mathbf{R}_{H_p\hat{H}_{LS}}^{-1} \hat{\mathbf{H}}_{LS} \\ &= \mathbf{R}_{H_pH_p} \left(\mathbf{R}_{H_pH_p} + \sigma_w^2 \left(\mathbf{X}_p \mathbf{X}_p^H \right)^{-1} \right)^{-1} \hat{\mathbf{H}}_{LS},\end{aligned}\quad (28)$$

where $\hat{\mathbf{H}}_{LS}$ is the LS estimate of \mathbf{H}_p derived in Equation 27, σ_w^2 is the common variance of W_k and w_n , and the covariance matrices are defined as follows:

$$\begin{aligned}\mathbf{R}_{H_pH_p} &= E\left\{ \mathbf{H}_p \mathbf{H}_p^H \right\}, \\ \mathbf{R}_{H_p\hat{H}_{LS}} &= E\left\{ \mathbf{H}_p \hat{\mathbf{H}}_{LS}^H \right\}, \\ \mathbf{R}_{\hat{H}_{LS}\hat{H}_{LS}} &= E\left\{ \hat{\mathbf{H}}_{LS} \hat{\mathbf{H}}_{LS}^H \right\}.\end{aligned}$$

It is observed from Equation 28 that a matrix inversion operation is involved in the MMSE estimator, and it must be calculated symbol by symbol. This problem can be solved by using a constant pilot, e.g., $X_{p,m} = c$, $m = 0, 1, \dots, N_p - 1$. A generic CE can be obtained by averaging over a sufficiently long duration of transmitted symbols (Edfors et al., 1996), i.e.,

$$\hat{\mathbf{H}}_{LMMSE} = \mathbf{R}_{H_pH_p} \left(\mathbf{R}_{H_pH_p} + \frac{\beta}{\Gamma} \mathbf{I} \right)^{-1} \hat{\mathbf{H}}_{LS}, \quad (29)$$

where $\Gamma = \frac{E\{|X_{p,k}|^2\}}{\sigma_w^2}$ is the average signal-to-noise ratio (SNR) and $\beta = E\{|X_{p,k}|^2\}E\{|1/X_{p,k}|^2\}$ is a constant determined by the signal mapping method employed in the pilot symbols. For example, $\beta = 17/9$ if 16-QAM is employed in the pilot symbols. If the auto-correlation matrix $\mathbf{R}_{H_p H_p}$ and the value of the SNR are known in advance, $\mathbf{R}_{H_p H_p} \left(\mathbf{R}_{H_p H_p} + \frac{\beta}{\Gamma} \mathbf{I} \right)^{-1}$ only needs to be calculated once. As shown in Equation 29, the CE requires N_p complex multiplications per pilot sub-carrier. To further reduce the number of multiplication operations, a low-rank approximation method based on singular-value decomposition (SVD) was adopted in the previous study (Edfors et al., 1996). Initially, the channel correlation matrix can be decomposed as

$$\mathbf{R}_{H_p H_p} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H, \quad (30)$$

where \mathbf{U} is a matrix with orthonormal columns $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{N_p-1}$, and $\mathbf{\Lambda}$ is a diagonal matrix with singular values $\lambda_0, \lambda_1, \dots, \lambda_{N_p-1}$ as its diagonal elements. The rank- ϱ approximation of the LMMSE CE derived in Equation 29 can thus be formulated as

$$\hat{\mathbf{H}}_{SVD} = \mathbf{U} \begin{bmatrix} \Delta_\varrho & 0 \\ 0 & 0 \end{bmatrix} \mathbf{U}^H \hat{\mathbf{H}}_{LS}, \quad (31)$$

where Δ_ϱ denotes a diagonal matrix with terms that can be expressed as

$$\delta_k = \frac{\lambda_k}{\lambda_k + \frac{\beta}{\Gamma}}, \quad k = 0, 1, \dots, \varrho. \quad (32)$$

After some manipulation, the CE in Equation 31 requires $2\varrho N_p$ complex multiplications, and the total number of multiplications per pilot tone becomes 2ϱ . In general, the number of essential singular values, ϱ , is much smaller than the number of pilot sub-channels, N_p , and the computational complexity is therefore considerably reduced when the low-rank SVD-based CE is compared with the full-rank LMMSE-based CE derived in Equation 29. Incidentally, low-rank SVD-based CE can combat parameter mismatch problems, as shown in previous studies (Edfors et al., 1996).

3.3 Data sub-channel interpolation

After joint estimation of the FD channel weights from the pilot sub-channels is complete, the channel weight estimation on the data (non-pilot) sub-channels must be interpolated from the pilot sub-channel estimates. A piecewise-linear interpolation method has been studied (Rinne & Renfors, 1996) that exhibits better CE performance than piecewise-constant interpolation. A piecewise-linear interpolation (LI) method, a piecewise second-order polynomial interpolation (SOPI) method and a transform-domain interpolation method are studied in this sub-section.

3.3.1 Linear interpolation

In the linear interpolation method, the channel weight estimates on any two adjacent pilot sub-channels are employed to determine the channel weight estimates of the data sub-channel located between the two pilot sub-channels (Rinne & Renfors, 1996). The channel estimate of the k th data sub-channel can be obtained by the LI method, i.e.,

$$\hat{H}_{LI,x,k} = \hat{H}_{LI,x,mQ+l} = \left(1 - \frac{l}{Q}\right) \hat{H}_{x,m} + \frac{l}{Q} \hat{H}_{x,m+1}, \quad \begin{array}{l} x = LS, LMMSE, SVD, \\ m = 0, 1, \dots, N_p - 2, \\ 1 \leq l \leq (Q - 1), \end{array} \quad (33)$$

where $mQ < k = mQ + l < (m + 1)Q$, $m = \lfloor \frac{k}{Q} \rfloor$, $\lfloor \cdot \rfloor$ denotes the greatest integer less than or equal to the argument and l is the value of k modulo Q .

3.3.2 Second-order polynomial interpolation

Intuitively, a higher-order polynomial interpolation may fit the CTF better than the aforementioned first-order polynomial interpolation (LI). The SOPI can be implemented with a linear, time-invariant FIR filter (Liu & Wei, 1992), and the interpolation can be written as

$$\begin{aligned} \hat{H}_{SOPI,k} &= \hat{H}_{SOPI,x,mQ+l} \\ &= c_1 \hat{H}_{x,m-1} + c_0 \hat{H}_{x,m} + c_{-1} \hat{H}_{x,m+1}, \end{aligned} \quad (34)$$

where

$$\begin{aligned} x &= LS, LMMSE, SVD, & m &= 1, 2, \dots, N_p - 2, & 1 \leq l \leq (Q - 1), \\ c_1 &= \frac{\psi(\psi + 1)}{2}, & c_0 &= -(\psi - 1)(\psi + 1), & c_{-1} &= \frac{\psi(\psi - 1)}{2}, \\ \psi &= \frac{l}{N}. \end{aligned}$$

3.3.3 Transform-domain-processing-based interpolation (TFDI)

An ideal low-pass filtering method based on transform-domain processing was adopted for the data sub-channel interpolation (Zhao & Huang, 1997). In accordance with the CTPA, the pilot sub-channels are equally spaced every Q sub-channels. This implies that the coherence bandwidth of the multipath fading channel under consideration is sufficiently wider than the bandwidth occupied by Q sub-channels. After the pilot sub-channel estimation was completed, the interpolation methods mentioned in 3.3.2 and 3.3.3 were used to search for some low-order-polynomial-based estimations (say, LI and SOPI) of the channel weights of the data sub-channels. A transform-domain-processing-based interpolation (TFDI) method proposed in a previous study was used to jointly smooth/filter out the sub-channel weight estimates of the data sub-channels (Zhao & Huang, 1997). The TFDI method consists of the following steps: (1) first, it transforms the sub-channel weight estimates obtained from the pilot sub-channels into the transform domain, which can be thought of as the TD here; (2) it keeps the essential elements unchanged, which include at most the leading N_p (multipath) components because the coherence bandwidth is as wide as N/N_p sub-channels; (3) it sets the tail $(N - N_p)$ components to zero; and (4) finally, it performs the inverse transformation

back to the sub-carrier domain, which may be called the FD in other publications. In this approach, a high-resolution interpolation method based on zero-padding and DFT/IDFT (Elliott, 1988) is employed. The TFDI technique can be thought of as ideal interpolation using an ideal lowpass filter in the transform domain.

3.4 Remarks

In this section, FD CE techniques based on CTPA were studied. Pilot sub-channel estimation techniques based on LS, LMMSE and SVD methods were studied along with data sub-channel interpolation techniques based on LI, SOPI and TFDI. The material provided in this section may also be found in greater detail in many prior publications (cited in this section) for interested readers. Of course, this author strongly encourages potential readers to delve into relevant research.

Many previous studies, e.g., (Zhao & Huang, 1997; Hsieh & Wei, 1998), prefer to adopt the IDFT as $\sum_{k=0}^{N-1} X_k e^{j2\pi kn/N}$ and the DFT as $\frac{1}{N} \sum_{k=0}^{N-1} x_n e^{j2\pi kn/N}$, rather than adopt those written in Equations 16 and 20. Although these representations are equivalent from the viewpoint of signal power, the formulations in Equations 16 and 20 are definitely more effective and convenient because they can keep the post-DFT-demodulation noise variance the same as the pre-DFT-demodulation noise variance. While performance analysis or comparison is conducted in terms of SNR, readers should be noted to take much care on this issue.

4. Time-domain channel estimation based on least-squares technique

4.1 Preliminary

A LS CE technique for mobile OFDM communication over a rapidly time-varying frequency-selective fading channel is demonstrated in this section. The studied technique, which uses CTPA, achieves low error probabilities by accurately estimating the CIR and effectively tracking rapid CIR time-variations. Unlike the technique studied in Section 3, the LS CE technique studied in this section is conducted in the TD, and several virtual sub-carriers are used. A generic estimator is performed serially block by block without assistance from a priori channel information and without increasing the computational complexity. The technique investigated in this section is also resistant to residual timing errors that occur during DFT demodulation. The material studied in this section has been thoroughly documented in a previous study and its references (Lin, 2008c). The author strongly encourages interested readers to look at these previous publications to achieve a deeper and more complete understanding of the material.

4.2 System description

The base-band signal $\{x_n\}$ consists of $2K$ complex sinusoids, which are individually modulated by $2K$ complex information-bearing QPSK symbols $\{X_k\}$, i.e.,

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=-K}^{K-1} X_{|k|_N} e^{j2\pi nk/N}, \quad n = 0, 1, \dots, N-1, \quad N \geq 2K, \quad (35)$$

where $X_{|k|_N}$ denotes the complex symbol transmitted on the $|k|_N$ th sub-channel, N is the IDFT size, n is the TD symbol index, k is the FD subcarrier index, $2K$ is the total number of

sub-channels used to transmit information and $|k|_N$ denotes the value of k modulo N . In Equation 35, $N_v = N - 2K$ sub-carriers are appended in the high-frequency bands as virtual sub-carriers and can be considered to be guard bands that avoid interference from other applications in adjacent bands and are not employed to deliver any information. It should be noted that x_n and X_k form an N -point DFT pair, i.e., $\text{DFT}_N \{x_n, n = 0, 1, \dots, N-1\} = \{X_0, X_1, \dots, X_{K-1}, 0, 0, \dots, 0, X_{N-K}, X_{N-K+1}, \dots, X_{N-1}\}$, where the 0s denote the symbols transmitted via the virtual sub-channels. In a CTPA OFDM system, the symbols transmitted on the sub-channels can be expressed in vector form for simplicity:

$$\mathbf{X} = \{X_k\} \in \mathbf{C}^{N \times 1}, \quad (36)$$

where

$$X_k = \begin{cases} 0, & k \in \zeta_v = \{K, K+1, \dots, N-K-1\} \\ P_l, & k \in \zeta_p = \{|(N-K) + (Q+1)/2 + l \cdot Q|_N | l = 0, 1, \dots, N_p-1\} \\ D_{k'}, & k \in \zeta \setminus \zeta_v \setminus \zeta_p; \end{cases}$$

$\zeta = \{0, 1, \dots, N-1\}$; $K = (N - N_v)/2$; P_l denotes the l th pilot symbol; N_p denotes the number of pilot sub-channels; Q denotes the pilot sub-channel separation, which is an odd number in the case under study; $D_{k'}$ represents the k' th information-bearing data symbol; ζ_v stands for the set of indices of the virtual sub-channels; and ζ_p stands for the set of indices of the pilot sub-channels. The OFDM block modulation can be reformulated as the following matrix operation:

$$\mathbf{x} = \mathbf{F}_1 \cdot \mathbf{X}, \quad (37)$$

where

$$\begin{aligned} \mathbf{x} &= \{x_n\} \in \mathbf{C}^{N \times 1}; \\ \mathbf{F}_1 &= \{f_{n,k}\} \in \mathbf{C}^{N \times N}, \\ f_{n,k} &= \frac{1}{\sqrt{N}} \exp\left(j \frac{2\pi kn}{N}\right), \quad 0 \leq k \leq N-1, \quad 0 \leq n \leq N-1. \end{aligned}$$

\mathbf{x} in Equation 37 can be rewritten as follows:

$$\mathbf{x} = \bar{\mathbf{x}} + \tilde{\mathbf{x}}, \quad (38)$$

where

$$\begin{aligned} \bar{\mathbf{x}} &= \{\bar{x}_n\} \in \mathbf{C}^{N \times 1}, \\ \bar{x}_n &= \frac{1}{\sqrt{N}} \sum_{l=0}^{N_p-1} P_l \cdot \exp\left(j 2\pi n \left| (N-K) + (Q+1)/2 + l \cdot Q \right|_N / N\right), \\ n &= 0, 1, \dots, N-1, \end{aligned}$$

which is the TD sequence obtained from the pilot symbols modulated on the pilot sub-channels; and

$$\begin{aligned}\tilde{\mathbf{x}} &= \{\tilde{x}_n\} \in \mathbf{C}^{N \times 1}, \\ \tilde{x}_n &= \frac{1}{\sqrt{N}} \sum_{\substack{k=0 \\ k \notin \zeta_v \\ k \notin \zeta_p}}^{N-1} X_k \cdot \exp(j2\pi nk / N), \quad n=0,1,\dots,N-1,\end{aligned}$$

which is the TD sequence that results from the information-bearing QPSK symbols modulated on the data (non-pilot) sub-channels. In accordance with the CLT, \tilde{x}_n , $n=0,1,\dots,N-1$ are independent, identically distributed (IID) zero-mean Gaussian random variables with variance $\frac{N-N_v-N_p}{N}\sigma_{X_k}^2$, where $\sigma_{X_k}^2$ is the transmitted signal power.

After the TD signal \mathbf{x} is obtained by conducting the IDFT modulation, a CP with length L is inserted, and the resulting complex base-band transmitted signal \mathbf{s} can be expressed as

$$\mathbf{s} = \mathbf{G}_I \cdot \mathbf{x} = \mathbf{G}_I \cdot \mathbf{F}_I \cdot \mathbf{X} \in \mathbf{C}^{(N+L) \times 1}, \quad (39)$$

where

$$\mathbf{G}_I = \begin{bmatrix} \mathbf{0}_{L \times (N-L)} & \mathbf{I}_L \\ & \mathbf{I}_N \end{bmatrix} \in \mathbf{N}^{(N+L) \times N},$$

\mathbf{G}_I is the matrix for CP insertion, \mathbf{I} is an identity matrix of the size noted in the subscript and $\mathbf{0}$ is a matrix of the size noted in the subscript whose entries are all zeros. The transmitted signal \mathbf{s} is fed into a parallel-to-serial (P/S) operator, a digital-to-analog converter (DAC), a symbol shaping filter and finally an RF modulator for transmission. For complex base-band signals, the equivalent base-band representation of a multipath channel can be expressed as $\hat{h}(\tau, t) = \sum_{m'} h_{m'}(t) \delta(\tau - \tau_{m'})$, where t denotes the time parameter, $h_{m'}(t)$ represents the m' th tap-weighting coefficient and τ is the delay parameter. The above 2-parameter channel model obeys the wide-sense stationary uncorrelated scattering (WSSUS) assumption. Based on the WSSUS and quasi-stationary assumptions, the channel tap-weighting coefficients are time-varying but do not change significantly within a single OFDM block duration of length NT_s , where T_s is the sampling period. Because the fractional durations (i.e., in a fraction of T_s) of delays are not taken into consideration, for a given time instant the above-mentioned tapped-delay-line channel model can be thought of as a CIR. Therefore, the channel model can be rewritten in a discrete-time representation for simplicity as $\mathbf{h} = \{h_m\} \in \mathbf{C}^{M \times 1}$, where M depends on the multipath delay spread. MT_s is thus the longest path delay; M varies according to the operating environment and cannot be known a priori at the receiving end. The received OFDM symbols can then be written in the following vector representation: $\mathbf{r}' = \mathbf{s} * \mathbf{h} + \mathbf{w}'$, where $*$ denotes the convolution operation, $\mathbf{r}' \in \mathbf{C}^{(N+L+M-1) \times 1}$ and \mathbf{w}' is an AWGN vector whose elements are IID zero-mean Gaussian random variables with variance σ_w^2 . While in practice a residual timing error \mathcal{G} may occur with the employed symbol timing synchronization mechanism, the steady-state-response portion of \mathbf{r}' can hopefully be obtained from

$$\mathbf{r}_g = \mathbf{G}_{R,\mathcal{G}} \cdot \mathbf{r}', \quad \mathbf{G}_{R,\mathcal{G}} = \begin{bmatrix} \mathbf{0}_{N \times (L-\mathcal{G})} & \mathbf{I}_N & \mathbf{0}_{N \times (M+\mathcal{G}-1)} \end{bmatrix}. \quad (40)$$

If the residual timing error \mathcal{G} in the above equation falls within $[0, L - M]$, there is no ISI in the received signal. In practice, \mathcal{G} may be only a few samples long and may be less than M , and $\mathcal{G} = 0$ represents perfect synchronization. The demodulation process at the receiving end can be performed by means of a DFT operation, and the received signal vector should thus be transformed back into the sub-carrier space, i.e.,

$$\mathbf{R}_g = \mathbf{F}_T \cdot \mathbf{r}_g \in \mathbf{C}^{N \times 1}, \quad (41)$$

where

$$\begin{aligned} \mathbf{F}_T &= \{f'_{k,n}\} \in \mathbf{C}^{N \times N}, \\ f'_{k,n} &= \frac{1}{\sqrt{N}} \exp(-j2\pi kn / N), \quad n = 0, 1, \dots, N-1; \quad k = 0, 1, \dots, N-1. \end{aligned}$$

Moreover, \mathbf{F}_T is the complex conjugate of \mathbf{F}_I defined below Equation 37 and denotes the DFT matrix. Thus, the demodulated signals \mathbf{R}_g on the sub-channels are obtained by the DFT operation, as shown in Equation 41. In addition, some specific components of \mathbf{R}_g represent the outputs of the transmitted pilot symbols that pass through the corresponding pilot sub-channels. These entries of \mathbf{R}_g , i.e., R_k^g , $k \in \zeta_p$, are exploited to estimate the pilot sub-channel by FDLS estimation, LMMSE or a complexity-reduced LMMSE via SVD, as shown in the previous section. After the pilot sub-channel gains have been estimated by FDLS, LMMSE or SVD, smoothing or interpolation/extrapolation methods are used to filter out the estimates of the data sub-channel gains from inter-path interference (IPI), ICI and noise. The previously mentioned pilot sub-channel estimation and data sub-channel interpolation/extrapolation can often be considered to be an up-sampling process conducted in the FD and can therefore be performed fully on the sub-channel space studied in Section 3.

As a matter of fact, the studied technique exploits a TD LS (TDLS) method to estimate the leading channel tap-weighting coefficients in the CIR, performs zero-padding to form an N -element vector and finally conducts the DFT operation on the resultant vector to effectively smooth in the FD. The studied technique accomplishes ideal interpolation with the domain transformation method used previously (Zhao & Huang, 1997). The whole CTF, including all of the channel gains on the pilot, data and virtual sub-channels over the entire occupied frequency band, can therefore be estimated simultaneously. The multipath delay spread of the transmission channel is typically dynamic and cannot be determined a priori at the receiving end. Therefore, the number of channel tap-weighting coefficients is often assumed to be less than L to account for the worst ISI-free case. The training sequence $\bar{\mathbf{x}}$ in the time direction, which is actually IDFT-transformed from the N_p in-band pilot symbols, has a period of approximately $\frac{N}{Q}$ because the pilot sub-channels are equally spaced by Q sub-channels. Therefore, the studied technique based on CTPA can effectively estimate at most the leading $\frac{N}{Q}$ channel tap-weighting coefficients. Meanwhile, in accordance with the Karhunen-Loeve (KL) expansion theorem (Stark & Woods, 2001), the training sequence $\bar{\mathbf{x}}$ can be considered to be a random sequence with N_p degrees of freedom. Therefore, the order of the TDLS technique studied in this section can be conservatively determined to be at most N_p because $\bar{\mathbf{x}}$ can be exploited to sound a channel with an order less than or equal to N_p . Based on the above reasoning, the number of channel tap-weighting coefficients is assumed to be less than or equal to N_p , and the longest excess delay is thus assumed to be less than $N_p T_s$. Therefore, the received signals \mathbf{r}_g can be reformulated as

$$\mathbf{r}_g = \mathbf{c}_g \cdot \mathbf{g} + \mathbf{w}_g = \bar{\mathbf{c}}_g \cdot \mathbf{g} + \tilde{\mathbf{w}}_g, \quad (42)$$

where $\mathbf{c}_g = \bar{\mathbf{c}}_g + \tilde{\mathbf{c}}_g$; $\tilde{\mathbf{w}}_g = \tilde{\mathbf{c}}_g \cdot \mathbf{g} + \mathbf{w}_g$;

$$\begin{aligned} \mathbf{c}_g &= \left\{ c_{p,q}^g \right\} \in \mathbf{C}^{N \times N_p}, & c_{p,q}^g &= x_{|p-g-q|_N} = \bar{x}_{|p-g-q|_N} + \tilde{x}_{|p-g-q|_N}, \\ \bar{\mathbf{c}}_g &= \left\{ \bar{c}_{p,q}^g \right\} \in \mathbf{C}^{N \times N_p}, & \bar{c}_{p,q}^g &= \bar{x}_{|p-g-q|_N}, \\ \tilde{\mathbf{c}}_g &= \left\{ \tilde{c}_{p,q}^g \right\} \in \mathbf{C}^{N \times N_p}, & \tilde{c}_{p,q}^g &= \tilde{x}_{|p-g-q|_N}, \\ & & 0 \leq p \leq N-1, & \quad 0 \leq q \leq N_p-1; \end{aligned}$$

\mathbf{c}_g is an $N \times N_p$ circulant matrix, and its left-most column is represented by

$$\text{column}_0(\mathbf{c}_g) = \left[x_{|N-g|_N} \quad x_{|N+1-g|_N} \quad \cdots \quad x_{|N-1-g|_N} \right]^T;$$

$\mathbf{w}_g = \{w_{k-g}\} \in \mathbf{C}^{N \times 1}$ is an AWGN vector whose N elements, w_{k-g} , $k = L, L+1, \dots, L+N-1$, are IID zero-mean Gaussian random variables with variance σ_w^2 ; and $\mathbf{g} = \{g_m\} \in \mathbf{C}^{N_p \times 1}$ contains the effective components that represent the channel tap-weighting coefficients. If no residual timing error exists, i.e., $g = 0$, then $g_m = h_m$, $m = 0, 1, \dots, M-1$ and $g_m = 0$, $M \leq m < N_p$. Here $M \leq N_p$, and at least $(N_p - M)$ components in \mathbf{g} must be zeros due to the lack of precise information about M at the receiving end, especially given that mobile OFDM communication systems often operate on a rapidly time-varying channel. As a result, the CIR can be estimated by means of a standard over-determined LS method, i.e.,

$$\hat{\mathbf{g}}^{\text{TDLS}} = \left\{ \hat{g}_m^{\text{TDLS}} \right\} = \left(\bar{\mathbf{c}}_0^H \bar{\mathbf{c}}_0 \right)^{-1} \bar{\mathbf{c}}_0^H \cdot \mathbf{r}_g \in \mathbf{C}^{N_p \times 1}, \quad (43)$$

where the superscript $(\cdot)^H$ denotes a Hermitian operator, and

$$\bar{\mathbf{c}}_0 = \left\{ \bar{c}_{p,q}^0 \right\} \in \mathbf{C}^{N \times N_p}, \quad \bar{c}_{p,q}^0 = \bar{x}_{|p-q|_N}, \quad 0 \leq p \leq N-1, \quad 0 \leq q \leq N_p-1.$$

In practice, a residual timing error that occurs in the DFT demodulation process inherently leads to phase errors in rotating the demodulated symbols. The phase errors caused by a timing error g are linearly dependent on both the timing error g and the sub-channel index k . Any small residual timing error can severely degrade the transmission performance in all of the previous studies that exploit two-stage CTPA CEs (Hsieh & Wei, 1998; Edfors et al., 1998; Seller, 2004; Edfors et al., 1996; Van de Beek et al., 1995; Park et al., 2004; Zhao & Huang, 1997). On the other hand, the studied technique has a higher level of tolerance to timing errors. Because the timing error g that occurs with the received training sequence (i.e., delayed replica of $\bar{\mathbf{x}}$) is the same as the error that occurs with the received data sequence (i.e., delayed replica of $\tilde{\mathbf{x}}$), the extra phase errors inserted into the demodulated symbols on individual sub-channels are the same as those that occur in the estimates of the sub-channel gains. Therefore, the extra phase rotations in the studied technique can be completely removed in the succeeding single-tap equalization process conducted on individual sub-channels. As a result, the studied technique can effectively deal with the problems caused by a residual timing error.

4.3 Remarks

The TD LS CE technique for OFDM communications has been studied in practical mobile environments. The studied TDLS technique based on the CTPA can accurately estimate the CIR and effectively track rapid CIR variations and can therefore achieve low error probabilities. A generic estimator is also performed sequentially on all OFDM blocks without assistance from a priori channel information and without increasing the computational complexity. Furthermore, the studied technique also exhibits better robustness to residual timing errors that occur in the DFT demodulation.

Whether OFDM communication should employ FD CE or TD CE has become an endless debate, because FD CE and equalization have attracted significant attention in recent years. While the LS method is not new, the TD CE may also not be considered novel. Although authors of some other publications thought that TDLS CE was not important, this must be a misunderstanding, and this section provides a very practical study. The material studied in this section has been deeply investigated in a previous study and its references (Lin, 2008c). This author strongly encourages interested readers, especially practical engineers and potential researchers, to examine the study and references closely to gain a deeper understanding of the applicability and practical value of the OFDM TD LS CE.

5. Channel estimation based on block pilot arrangement

5.1 Preliminary

The preceding two sections describe CE techniques based on the CTPA and taking advantage of either FD estimation or TD estimation methods. A CE technique based on the BTPA is discussed in this section. SC-FDMA has been chosen in the LTE specifications as a promising uplink transmission technique because of its low PAPR. Moreover, SC-FDM systems can be considered to be pre-coded OFDM communication systems, whose information symbols are pre-coded by the DFT before being fed into a conventional OFDMA (Myung et al., 2006).

In practice, pilot signals or reference signals for CE in SC-FDMA systems are inserted to occupy whole sub-channels periodically in the time direction, which can be considered to be BTPA. In this section, the signal model and system description of a BTPA-based CE technique is studied. The material discussed in this section can be found, in part, in a previous study (Huang & Lin, 2010).

5.2 System description

The information-bearing Gray-encoded symbols $\chi_u[n]$, $n = 0, 1, \dots, N_u - 1$ are pre-spread by an N_u -point DFT to generate the FD symbols $X_u[\kappa]$, $\kappa = 0, 1, \dots, N_u - 1$, i.e.,

$$X_u[\kappa] = \frac{1}{\sqrt{N_u}} \sum_{n=0}^{N_u-1} \chi_u[n] e^{-j2\pi n\kappa/N_u}, \quad \begin{array}{l} \kappa = 0, 1, \dots, N_u - 1, \\ u = 0, 1, \dots, U - 1, \end{array} \quad (44)$$

where U denotes the number of the users transmitting information toward the base-station, u denotes the user index, N_u denotes the sub-channel number which the u th user occupies, n denotes the time index and κ denotes the sub-carrier index. For a localized chunk arrangement used in the LTE specification, $X_u[\kappa]$, $\kappa = 0, 1, \dots, N_u - 1$ are allocated onto N_u sub-channels, i.e.,

$$S_u[k] = \begin{cases} X_u[k], & k = \Gamma_u(\kappa) = \sum_{i=0}^{u-1} N_i + \kappa \\ 0, & k \neq \Gamma_u(\kappa), \quad \kappa = 0, 1, \dots, N_u - 1. \end{cases} \quad (45)$$

The transmitted signal of the u th user is given by

$$s_u[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} S_u[k] e^{j2\pi kn/N}, \quad \begin{matrix} n = 0, 1, \dots, N-1; \\ u = 0, 1, \dots, U-1. \end{matrix} \quad (46)$$

The signal received at the base-station can be expressed as

$$r[n] = \sum_{u=0}^{U-1} \sum_{m=0}^{M-1} h_u[m, n] s_u[n-m] + w[n], \quad n = 0, 1, \dots, N-1, \quad (47)$$

where $h_u[m, n]$ is the sample-spaced channel impulse response of the m th resolvable path on the time index n for the u th user, M denotes the total number of resolvable paths on the frequency-selective fading channel and $w[n]$ is AWGN with zero mean and a variance of σ_w^2 . The time-varying multipath fading channel considered here meets the WSSUS assumption. Therefore, the channel-weighting coefficient $h_u[m, n]$ is modelled as a zero-mean complex Gaussian random variable, with an autocorrelation function that is written as

$$\mathbb{E}\{h_u[m, n] h_u^*[k, l]\} = \sigma_u^2[m] J_0(2\pi v_u |n-l| T_s) \delta[m-k], \quad (48)$$

where $\delta[\cdot]$ denotes the Dirac delta function, $J_0(\cdot)$ denotes the zeroth-order Bessel function of the first kind, v_u denotes the maximum Doppler frequency of the u th user and $\sigma_u^2[m]$ denotes the power of the m th resolvable path on the channel that the u th user experiences. In addition, it is assumed in the above equation that the channel tap-weighting coefficients on different resolvable paths are uncorrelated and that the channel tap-weighting coefficients on an individual resolvable path have the Clarke's Doppler power spectral density derived by Jakes (Jakes & Cox, 1994). To simplify the formulation of Equation 47, it is assumed that timing synchronization is perfect, ISI can be avoided and CP can be removed. At the receiving end, the FFT demodulation is conducted, and the received TD signal $r[n]$ is thus transformed into the FD for demultiplexing, i.e.,

$$\begin{aligned} R[k] &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} r[n] e^{-j2\pi nk/N} \\ &= \sum_{u=0}^{U-1} \bar{H}_{u, N/2}[k] S_u[k] + W[k], \quad k = 0, 1, \dots, N-1, \end{aligned} \quad (49)$$

where

$$\begin{aligned} \bar{H}_{u, N/2}[k] &\doteq H_u[k, n], & n = 0, 1, \dots, N-1, \\ H_u[k, n] &= \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} h_u[m, n] e^{-j2\pi mk/N}, & \forall k = \Gamma_u(\kappa), \\ W[k] &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} w[n] e^{-j2\pi nk/N}, & k = 0, 1, \dots, N-1. \end{aligned}$$

In conventional FD CE, the weighting coefficient on the k th sub-channel $\bar{H}_{u,N/2}[k]$ is estimated by the FD LS CE, i.e.,

$$\hat{H}_{\text{FDLS}}[k] = \frac{R[k]S_p^*[k]}{|S_p[k]|^2}, \quad k = \Gamma_u(\kappa), \quad (50)$$

where $S_p[k]$ represents the pilot symbols in the FD, which are known a priori at the receiving end. In the LTE uplink, $S_p[k], \forall k$ are obtained by transforming a Zadoff-Chu sequence onto the sub-carrier domain. Several CE techniques have been discussed in greater detail in a previous study (Huang & Lin, 2010). When the CE conducted by taking advantage of the pilot block is complete, several interpolation (or extrapolation) methods are conducted in the time direction to effectively smooth (or predict) the CTF or CIR upon transmission of the information-bearing symbols.

6. Channel estimation assisted from time-domain redundancy

6.1 Preliminary

To illustrate CE assisted by TD redundancy, a LS CE technique is studied in this section. The studied technique can apply pseudo-random-postfix orthogonal-frequency-division multiplexing (PRP-OFDM) communications to mobile applications, which often operate on a rapidly time-varying frequency-selective fading channel. Because conventional techniques that exploit a moving-average filter cannot function on a rapid time-varying channel, the studied technique takes advantage of several self-interference cancellation (SIC) methods to reduce IPI, ISI and IBI effectively and in a timely manner. The studied technique can thus overcome frequency selectivity caused by multipath fading and time selectivity caused by mobility; in particular, OFDM communication is often anticipated to operate in environments where both wide Doppler spreads and long delay spreads exist. Because conventional techniques based on MMSE CE usually require a priori channel information or significant training data, the studied method exploits a generic estimator assisted by LS CE that can be performed serially, block by block, to reduce computational complexity.

6.2 System description

The i th $N \times 1$ digital input vector $\mathbf{X}_N[i]$ is first modulated at the transmitting end with an IDFT operation. Thus, the TD information-bearing signal block can be expressed as

$$\mathbf{x}'_N[i] = \mathbf{F}_N^H \mathbf{X}_N[i], \quad (51)$$

where $\mathbf{X}_N[i]$ contains $2K \leq N$ QPSK-mapping information-bearing symbols;

$$\mathbf{F}_N = \frac{1}{\sqrt{N}} \{W_N^{kl}\}, \quad W_N = e^{-j2\pi/N}, \quad 0 \leq k < N, \quad 0 \leq l < N.$$

Immediately after the IDFT modulation process, a postfix vector $\mathbf{c}'_L = [c_0 \ c_1 \ \dots \ c_{L-1}]^T$ is appended to the IDFT modulation output vector $\mathbf{x}'_N[i]$. In this section, \mathbf{c}'_L is sifted from a partial period of a long pseudo-random sequence, and \mathbf{c}'_L is phase-updated at every frame that contains several TD OFDM signal blocks, rather than using a deterministic postfix vector with a pseudo-random weight as in the conventional PRP-OFDM (Muck et al., 2006;

2005; 2003). This change is desirable when considering that previous works did not suggest long PRP sequences (Muck et al., 2006; 2005; 2003) and that pseudo-random sequences, e.g., the m-sequences or Gold sequences, are actually more general in various communication applications. Therefore, the i th transmitted block, with a length of $\Xi = N + L$, can be expressed as

$$\mathbf{x}_{\Xi}[i] = \mathbf{F}_{zp}^H \mathbf{X}_N[i] + \mathbf{c}_{\Xi}, \quad (52)$$

where

$$\mathbf{F}_{zp}^H = \begin{bmatrix} \mathbf{I}_N \\ \mathbf{0}_{L \times N} \end{bmatrix} \mathbf{F}_N^H, \quad \mathbf{c}_{\Xi} = \begin{bmatrix} \mathbf{0}_{N \times 1} \\ \mathbf{c}'_{L \times 1} \end{bmatrix}_{\Xi \times 1},$$

\mathbf{I}_N denotes an $N \times N$ identity matrix and $\mathbf{0}_{L \times N}$ denotes a zero matrix of the size indicated in the subscript. The elements of $\mathbf{x}_{\Xi}[i]$ are then transmitted sequentially one by one (probably with transmit filtering or symbol shaping).

The channel studied here is modelled with a tapped-delay line of order $v - 1$, i.e., the impulse response of the investigated channel can be written as $\mathbf{h} = [h_0 \ h_1 \ \dots \ h_{v-1}]^T$. It is commonly assumed that the length of the postfix (or prefix) L is larger than the length of the channel impulse response v . Typically, the multipath delay spread of the transmission channel is dynamic and cannot be determined a priori at the receiving end. Therefore, the number of channel tap-weighting coefficients is often assumed to be up to L to consider the worst ISI-free case, i.e., $v = L$. Thus, the longest excess delay is vT_s , where T_s denotes the sample duration.

At the receiving end, the i th OFDM symbol block can be formulated as

$$\mathbf{r}_{\Xi}[i] = (\mathbf{h}_{\text{IBI},\Xi} + \mathbf{h}_{\text{ISI},\Xi}) \mathbf{x}_{\Xi}[i] + \mathbf{w}_{\Xi}[i], \quad (53)$$

where $\mathbf{h}_{\text{IBI},\Xi}$ is an $\Xi \times \Xi$ Toeplitz upper-triangular matrix in which the upper-most row is represented by

$$\text{row}_0(\mathbf{h}_{\text{IBI},\Xi}) = [0 \ \dots \ 0 \ h_{v-1} \ h_{v-2} \ \dots \ h_1],$$

$\mathbf{h}_{\text{ISI},\Xi}$ is an $\Xi \times \Xi$ Toeplitz lower-triangular matrix in which the left-most column is represented by

$$\text{column}_0(\mathbf{h}_{\text{ISI},\Xi}) = [h_0 \ h_1 \ \dots \ h_{v-1} \ 0 \ \dots \ 0]^T;$$

and $\mathbf{w}_{\Xi}[i]$ is the i th AWGN vector of elements with variance σ_w^2 .

6.2.1 Channel estimation

In this section, the CIR is considered to be time-varying, but not significantly changing within one or two OFDM blocks. The symbols employed here in the CE can be written as follows:

$$\mathbf{r}_{\text{CE},L+v-1}[i] = \begin{bmatrix} \langle \mathbf{r}_{\Xi}[i-1]_{N:\Xi-1} \rangle \\ \langle \mathbf{r}_{\Xi}[i]_{0:v-2} \rangle \end{bmatrix}_{(L+v-1) \times 1}, \quad (54)$$

where $\langle \mathbf{A} \rangle_{p,q}$ denotes either a column vector with elements arranged as $[A_p A_{p+1} \cdots A_q]^T$, sifted from a column vector \mathbf{A} , or a row vector with elements arranged as $[A_p A_{p+1} \cdots A_q]$, sifted from a row vector \mathbf{A} . In fact, $\mathbf{r}_{\text{CE},L+v-1}[i]$ can be reformulated in detail as follows:

$$\mathbf{r}_{\text{CE},L+v-1}[i] = \mathbf{C}[i]\mathbf{h} + \mathbf{w}''[i] = \mathbf{C}_o\mathbf{h} + \mathbf{w}''[i], \quad (55)$$

where $\mathbf{w}''[i]$ is an $(L + v - 1) \times 1$ AWGN vector of elements whose variances are σ_w^2 ;

$$\mathbf{C}[i] = (\mathbf{C}_L[i] + \mathbf{C}_o + \mathbf{C}_U[i]),$$

\mathbf{C}_o is an $(L + v - 1) \times v$ Toeplitz matrix in which the left-most column is represented by

$$\text{column}_0(\mathbf{C}_o) = [c_0 \ c_1 \ \cdots \ c_{L-1} \ 0 \ \cdots \ 0]^T;$$

$\mathbf{C}_U[i]$ is an $(L + v - 1) \times v$ upper-triangular Toeplitz matrix in which the upper-most row is represented by

$$\text{row}_0(\mathbf{C}_U[i]) = [0 \ \langle \mathbf{x}_\Xi[i-1] \rangle_{N-1:N-(v-1)}];$$

$\mathbf{C}_L[i]$ is an $(L + v - 1) \times v$ lower-triangular Toeplitz matrix in which the left-most column is represented by

$$\text{column}_0(\mathbf{C}_L[i]) = [\mathbf{0}_{1 \times L} \ \langle \mathbf{x}_\Xi[i] \rangle_{0:v-2}^T]^T;$$

and

$$\mathbf{w}''[i] = \mathbf{w}''[i] + \mathbf{C}_L[i]\mathbf{h} + \mathbf{C}_U[i]\mathbf{h}.$$

In the above equation, $\mathbf{C}_L[i]\mathbf{h}$ results in ISI extending from on-time symbols onto the CE. Meanwhile, $\mathbf{C}_U[i]\mathbf{h}$ leads to IBI extending from preceding symbols onto the CE. In accordance with the LS philosophy (Stark & Woods, 2001; Kay, 1993), the CE studied here can thus be formulated as

$$\hat{\mathbf{h}}_0[i] = (\mathbf{C}_o^H \mathbf{C}_o)^{-1} \mathbf{C}_o^H \bar{\mathbf{r}}_{\text{CE},L+v-1}[i], \quad (56)$$

where

$$\bar{\mathbf{r}}_{\text{CE},L+v-1}[i] = \frac{1}{2} (\mathbf{r}_{\text{CE},L+v-1}[i] + \mathbf{r}_{\text{CE},L+v-1}[i+1]).$$

In fact, the CE performed using $\bar{\mathbf{r}}_{\text{CE},L+v-1}[i]$ forces the channel estimator $\hat{\mathbf{h}}_0[i]$, derived in Equation 56, to effectively exploit the first-order statistics to conduct the TD LI as employed in a previous work (Ma et al., 2006). Because of the LS philosophy, the statistics of $\mathbf{w}''[i]$ need not be completely known prior to performing the CE and $(\mathbf{C}_o^H \mathbf{C}_o)^{-1} \mathbf{C}_o^H$ can be pre-calculated and pre-stored as a generic LS CE to reduce complexity. Furthermore, by taking advantage of decision-directed (DD) SIC, estimates of the CIR can be iteratively obtained by

$$\hat{\mathbf{h}}_1[i] = \{(\mathbf{C}_o^H \mathbf{C}_o)^{-1} \mathbf{C}_o^H\} \tilde{\mathbf{r}}_{\text{CE},L+v-1}[i], \quad (57)$$

where

$$\begin{aligned}\tilde{\mathbf{r}}_{\text{CE},L+v-1}[i] &= \frac{1}{2} \left(\mathbf{r}_{\text{CE},L+v-1}[i] - \hat{\mathbf{C}}_{\text{U}}[i-1] \hat{\mathbf{h}}_1[i-1] + \mathbf{r}_{\text{CE},L+v-1}[i+1] \right), \quad i \geq 1; \\ \hat{\mathbf{h}}_1[0] &= \hat{\mathbf{h}}_0[0]; \quad \hat{\mathbf{C}}_{\text{U}}[0] \hat{\mathbf{h}}_1[0] = \mathbf{0}_{(L+v-1) \times 1} \text{ (initialization);}\end{aligned}$$

and $\hat{\mathbf{C}}_{\text{U}}[i-1]$ denotes an $(L+v-1) \times L$ upper-triangular Toeplitz matrix in which the upper-most row is $[0 \ \langle \hat{\mathbf{x}}_{\Xi}[i-1] \rangle_{N-1:N-(v-1)}]$, which results from the DD symbols. Eventually, the estimates of sub-channel gains in individual frequency bins can be obtained by performing the DFT on the zero-padded replicas of either $\hat{\mathbf{h}}_0[i]$ or $\hat{\mathbf{h}}_1[i]$, i.e.,

$$\hat{\mathbf{H}}_k[i] = \mathbf{F}_N \begin{bmatrix} \mathbf{I}_v & \mathbf{0}_{v \times (N-v)} \\ \mathbf{0}_{(N-v) \times v} & \mathbf{0}_{(N-v) \times (N-v)} \end{bmatrix} \hat{\mathbf{h}}_k[i], \quad k = 0, 1. \quad (58)$$

6.2.2 Symbol recovery

For information detection at the receiving end, the i th information symbol within the DFT window can be obtained as

$$\mathbf{r}_{\text{SD},N}[i] = \langle \mathbf{r}_{\Xi}[i] \rangle_{0:N-1}, \quad (59)$$

and thus, its corresponding FD symbol is

$$\mathbf{R}_{\text{SD},0,N}[i] = \mathbf{F}_N \mathbf{r}_{\text{SD},N}[i]. \quad (60)$$

OLA: Based on the signal formatting in the PRP-OFDM communication under investigation, the ICI caused by various excess delays can be taken into account by modifying the signal symbol for signal detection to be

$$\mathbf{R}_{\text{SD},1,N}[i] = \mathbf{F}_N \left(\mathbf{r}_{\text{SD},N}[i] + \mathbf{r}_{\text{IClC},N}[i] \right), \quad (61)$$

where

$$\mathbf{r}_{\text{IClC},N}[i] = \begin{bmatrix} \langle \mathbf{r}_{\Xi}[i] \rangle_{N:N+v-2} \\ \mathbf{0}_{(N-v+1) \times 1} \end{bmatrix}$$

is exploited here for the purpose of ICI compensation. In fact, $\mathbf{R}_{\text{SD},1,N}[i]$ in Equation 61 can be considered to be a complexity-reduced variant modified from the method that was called the overlap-add (OLA) approach in previous studies (Muquest et al., 2002; Muck et al., 2003). It has been proven in previous studies (Muquet et al., 2000) that the OLA helps the ZP-OFDM achieve the same performance as the CP-OFDM because the OLA can reduce ICI by compensating for IPI and timing errors to maintain the orthogonality among sub-carriers, as in the CP-OFDM.

OLA with SIC: The conventional OLA mentioned above introduces some self-interference to the PRP-OFDM. Therefore, the self-interference occurring in the PRP-OFDM signal detection has to be eliminated. As a result, the signals fed into the detection can be formulated as

$$\mathbf{R}_{SD,2,k,N}[i] = \mathbf{F}_N \left(\mathbf{r}_{SD,N}[i] + \mathbf{r}_{ICIC,N}[i] - \hat{\mathbf{r}}_{prp,k}[i] \right), \quad k = 0, 1,$$

where

$$\hat{\mathbf{r}}_{prp,k}[i] = \tilde{\mathbf{C}}_o \hat{\mathbf{h}}_k[i-1], \quad k = 0, 1$$

and

$$\tilde{\mathbf{C}}_o = \left\langle \tilde{\mathbf{C}}'_o \right\rangle_{(v-1) \times v}$$

denotes a matrix containing the most upper-left $(v-1) \times v$ elements of the matrix $\tilde{\mathbf{C}}'_o$, which is a circulant matrix in which the left-most column is $[c_0 \ c_1 \ \cdots \ c_{L-1}]^T$.

6.3 Remarks

The LS CE technique has been thoroughly investigated in practical mobile environments. By taking advantage of SIC mechanisms, the studied technique can efficiently eliminate various interferences, accurately estimate the CIR, effectively track rapid CIR variations and, therefore, achieve low error probabilities. The studied technique can also achieve low bit error floors. The generic estimator assisted by LS CE can be performed sequentially on all OFDM blocks for complexity reduction without a priori channel information, which is required by conventional techniques based on MMSE. Several previous studies and their references regarding this topic are worth noting (Lin, 2009b;a; Lin & Lin, 2009; Lin, 2008b;a).

7. Summary

In this chapter, a variety of CE techniques on OFDM communications were investigated. This author does not attempt to present this topic in detail nor provide theoretical derivations and rigorous statistical analysis, though they are thought of as the most crucial for a journal publication. Insightful and reader-friendly descriptions are presented to attract readers of any level, including practicing communication engineers and beginning and professional researchers. All interested readers can easily find noteworthy materials in much greater detail from previous publications and the references cited in this chapter.

8. References

- 4MORE (2005). Eu-ist-4more project website, www.ist-4more.org.
- Bingham, J. A. C. (1990). Multicarrier modulation for data transmission: an idea whose time has come, *IEEE Communications Magazine* Vol. 28(No. 5): 5-14.
- Chang, R.W. (1966). Synthesis of band-limited orthogonal signals for multichannel data transmission, *Bell System Technical Journal* Vol. 45(No. 12): 1775-1796.
- Chow, P. S. (1993). *Bandwidth Optimized Digital Transmission Techniques for Spectrally Shaped Channels with Impulse Noise*, Ph. D. Dissertation, Stanford University, CA.
- Coleri, S., Ergen, M., Puri, A. & Bahai, A. (2002). Channel estimation techniques based on pilot arrangement in ofdm systems, *IEEE Transactions on Broadcasting* Vol. 48(No. 3): 223-229.

- Couasnon, T. D., Monnier, R. & Rault, J. B. (1994). Ofdm for digital tv broadcasting, *Signal Processing* Vol. 39(No. 1-2): 1-32.
- DAB (1995). Radio broadcasting systems; digital audio broadcasting (dab) to mobile, portable and fixed receivers, *European Telecommunications Standards ETS 300 401*, ETSI.
- Darlington, S. (1970). On digital single-sideband modulators, *IEEE Transactions on Circuit Theory* Vol. 17(No. 3): 409-414.
- DVB (1996). Digital broadcasting systems for television, sound and data services, *European Telecommunications Standards prETS 300 744*.
- Edfors, O., Sandell, M., van de Beek, J.-J., Wilson, S. K. & Borjesson, P. O. (1996). Ofdm channel estimation by singular value decomposition, *Proceedings of IEEE 46th Vehicular Technology Conference, 1996*, IEEE Vehicular Technology Society, Atlanta, GA, pp. 923-927.
- Edfors, O., Sandell, M., van de Beek, J.-J., Wilson, S. K. & Borjesson, P. O. (1998). Ofdm channel estimation by singular value decomposition, *IEEE Transactions on Communications* Vol. 46(No. 7): 931-939.
- Elliott, D. F. (1988). *Handbook of Digital Signal Processing: Engineering Applications*, Academic Press.
- Fazel, K. (1994). Performance of convolutionally coded cdma/ofdm in a frequency-time selective fading channel and its near-far resistance, *Proceedings of IEEE International Conference on Communications, 1994. ICC'94*, IEEE Communications Society, New Orleans, LA, pp. 1438 - 1442.
- Floch, B. L., Alard, M. & Berrou, C. (1995). Coded orthogonal frequency division multiplex [tv broadcasting], *Proceedings of the IEEE* Vol. 86(No. 6): 982-996.
- Gui, L., Li, Q., Liu, B., Zhang, W. & Zheng, C. (2009). Low complexity channel estimation method for tds-ofdm based chinese dtmb system, *IEEE Transactions Consumer Electronics* Vol. 55(No. 3): 1135-1140.
- Han, K.-Y., Lee, S.-W., Lim, J.-S. & Sung, K.-M. (2004). Channel estimation for ofdm with fast fading channels by modified kalman filter, *IEEE Transactions on Consumer Electronics* Vol. 50(No. 2): 443-449.
- Hara, S. & Prasad, R. (1997). Multicarrier modulation for data transmission: an idea whose time has come, *IEEE Communications Magazine* Vol. 35(No. 12): 126-133.
- Hoeher, P. (1991). Tcm on frequency-selective land-mobile fading channels, *Proceedings of International Workshop Digital Communications, Tirrenia, Italy*, pp. 317-328.
- Hoeher, P., Kaiser, S. & Robertson, P. (1997). Two-dimensional pilot-symbol-aided channel estimation by wiener filtering, *Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, IEEE Signal Processing Society, Munich, pp. 1845-1848.
- Hsieh, M.-H. & Wei, C.-H. (1998). Channel estimation for ofdm systems based on comb-type pilot arrangement in frequency selective fading channels, *IEEE Transactions on Consumer Electronics* Vol. 44(No. 1): 217-225.
- Huang, S.-C. & Lin, J.-C. (2010). Novel channel estimation techniques on sc-fdma uplink transmission, *Proceedings of 2010 IEEE Vehicular Technology Conference (VTC 2010-Spring)*, IEEE Vehicular Technology Society, Taipei, Taiwan.
- Jakes, W. C. & Cox, D. C. (1994). *Microwave Mobile Communications*, Wiley-IEEE Press.

- Kay, S. M. (1993). *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall; 1 edition.
- Klerer, M. (2005). Ieee 802.20 websites, grouper.ieee.org/groups/802/20/.
- Kondo, S. & Milstein, B. (1996). Performance of multicarrier ds cdma systems, *IEEE Transactions on Communications* Vol. 44(No. 2): 238-246.
- Levanon, N. & Mozeson, E. (2004). *Radar Signals*, Wiley-IEEE Press.
- Li, B., Xu, Y. & Choi, J. (2002). A study of channel estimation in ofdm systems, *Proceedings of 2002 IEEE 56th Vehicular Technology Conference, 2002 (VTC 2002-Fall)*, IEEE Vehicular Technology Society, Vancouver, Canada, pp. 894-898.
- Li, B., Xu, Y. & Choi, J. (2008). Channel estimation for lte uplink in high doppler spread, *Proceedings of Wireless Communications and Networking Conference, 2008 (WCNC 2008)*, IEEE Communications Society, Las Vegas, NV, pp. 1126-1130.
- Lin, J.-C. (2008a). Channel estimation assisted by postfixed pseudo-noise sequences padded with null samples for mobile ofdm communications, *Proceedings of IEEE Wireless Communications and Networking Conference, 2008 (WCNC 2008)*, IEEE Communications Society, Las Vegas, NV, pp. 846-851.
- Lin, J.-C. (2008b). Least-squares channel estimation assisted by self-interference cancellation for mobile prp-ofdm applications, *Proceedings of IEEE International Conference on Communications, 2008 (ICC '08)*, IEEE Communications Society, Beijing, China, pp. 578- 583.
- Lin, J.-C. (2008c). Least-squares channel estimation for mobile ofdm communication on time-varying frequency-selective fading channels, *IEEE Transactions on Vehicular Technology* Vol. 57(No. 6): 3538-3550.
- Lin, J.-C. (2009a). 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): 561-570.
- Lin, J.-C. (2009b). Least-squares channel estimation assisted by self-interference cancellation for mobile prp-ofdm applications, *IET Communications* Vol. 3(No. 12): 1907-1918.
- Lin, J.-C. & Lin, C.-S. (2009). Ls channel estimation assisted from chirp sequences in ofdm communications, *Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009 (Wireless VITAE 2009)*, IEEE ComSoc, ITS and VTS, Aalborg, Denmark, pp. 222-226.
- Liu, G.-S. & Wei, C.-H. (1992). A new variable fractional sample delay filter with nonlinear interpolation, *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* Vol. 39(No. 2): 123-126.
- Liu, G. & Zhang, J. (2007). Itd-dfe based channel estimation and equalization in tds-ofdm receivers, *IEEE Transactions Consumer Electronics* Vol. 53(No. 2): 304-309.
- LTE (2009). TS 36.211 (V8.5.0), *Physical Channels and Modulation*, 3GPP.
- Ma, Y., Yi, N. & Tafazolli, R. (2006). Channel estimation for prp-ofdm in slowly time-varying channel: first-order or second-order statistics?, *IEEE Signal Processing Letters* Vol. 13(No. 3): 129-132.
- Marks, R. B. (2008). Ieee 802.16 standard, www.ieee802.org/16/.
- Marti, B., Bernard, P., Lodge, N. & Schafer, R. (1993). European activities on digital television broadcasting - from company to cooperative projects, *EBU Technical Review* Vol. 256: 20-29. www.ebu.ch/en/technical/trev/trev_256-marti.pdf.

- MATRICE (2005). Eu-ist-matrice project website, www.ist-matrice.org .
- Minn, H. & Bhargava, V. K. (1999). Channel estimation for ofdm systems with transmitter diversity immobile wireless channels, *IEEE Journal on Selected Areas in Communications* Vol. 17(No. 3): 461–471.
- Minn, H. & Bhargava, V. K. (2000). An investigation into time-domain approach for ofdm channel estimation, *IEEE Transactions on Broadcasting* Vol. 46(No. 4): 240–248.
- Moeneclaey, M. & Bladel, M. V. (1993). Digital hdtv broadcasting over the catv distribution system, *Signal processing: Image communication* Vol. 5(No. 5-6): 405–415.
- Muck, M., de Courville, M., Debbah, M. & Duhamel, P. (2003). A pseudo random postfix ofdm modulator and inherent channel estimation techniques, *Proceedings of IEEE 2003 Global Telecommunications Conference (GLOBECOM '03)*, IEEE Communications Society, San Francisco, pp. 2380–2384.
- Muck, M., de Courville, M. & Duhamel, P. (2006). A pseudorandom postfix ofdm modulator - semi-blind channel estimation and equalization, *IEEE Transactions on Signal Processing* Vol. 54(No. 3): 1005–1017.
- Muck, M., de Courville, M., Miet, X. & Duhamel, P. (2005). Iterative interference suppression for pseudo random postfix ofdm based channel estimation, *Proceedings of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, IEEE Signal Processing Society, Philadelphia, Pennsylvania, USA, pp. 765–768.
- Muquest, B., Wang, Z., Giannakis, G. B., Courville, M. & Duhamel, P. (2002). Cyclic prefixing or zero padding for wireless multicarrier transmissions?, *IEEE Transactions on Communications* Vol. 50(No. 12): 2136–2148.
- Muquet, B., de Courville, M., Giannakis, G. B., Wang, Z. & Duhamel, P. (2000). Reduced-complexity equalizers for zero-padded ofdm transmissions, *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, IEEE Signal Processing Society, Istanbul, pp. 2973–2976.
- Myung, H. G., Lim, J. & Goodman, D. J. (2006). Single carrier fdma for uplink wireless transmission, *IEEE Vehicular Technology Magazine* Vol. 1(No. 3): 30–38.
- Negi, R. & Cioffi, J. (1998). Pilot tone selection for channel estimation in a mobile ofdm system, *IEEE Transactions on Consumer Electronics* Vol. 44(No. 3): 1122–1128.
- Ng, J. C. L., Letaief, K. B. & Murch, R. D. (1998). Complex optimal sequences with constant magnitude for fast channel estimation initialization, *IEEE Transactions on Communications* Vol. 46(No. 3): 305–308.
- Ohno, S. & Giannakis, G. B. (2002). Optimal training and redundant precoding for block transmissions with application to wireless ofdm, *IEEE Transactions on Communications* Vol. 50(No. 12): 2113–2123.
- Park, J., Kim, J., Kang, C. & Hong, D. (2004). Channel estimation performance analysis for comb-type pilot-aided ofdm systems with residual timing offset, *Proceedings of IEEE 60th Vehicular Technology Conference, 2004 (VTC2004-Fall)*, IEEE Vehicular Technology Society, Los Angeles, CA, pp. 4376–4379.
- Peled, A. & Ruiz, A. (1980). Frequency domain data transmission using reduced computational complexity algorithms, *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'80)*, IEEE Signal Processing Society, Denver, CO, pp. 964–967.

- Popovic, B. M. (1992). Generalized chirp-like polyphase sequences with optimum correlation properties, *IEEE Transactions on Information Theory* Vol. 38(No. 4): 1406-1409.
- Reiners, C. & Rohling, H. (1994). Multicarrier transmission technique in cellular mobile communication systems, *Proceedings of 1994 IEEE 44th Vehicular Technology Conference*, IEEE Vehicular Technology Society, Stockholm, pp. 1645-1649.
- Rinne, J. & Renfors, M. (1996). Optimal training and redundant precoding for block transmissions with application to wireless ofdm, *IEEE Transactions Consumer Electronics* Vol. 42(No. 4): 959-962.
- Saltzberg, B. (1967). Performance of an efficient parallel data transmission system, *IEEE Transactions on Communication Technology* Vol. 15(No. 6): 805-811.
- Sandell, M. & Edfors, O. (1996). A comparative study of pilot-based channel estimators for wireless ofdm, *Research Report / 1996:19. Div. Signal Processing, Lulea Univ. Technology, Lulea, Sweden* Vol.(No.).
- Seller, O. (2004). Low complexity 2d projection-based channel estimators for mc-cdma, *Proceedings of 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2004 (PIMRC 2004)*, IEEE Communications Society, Barselona, pp. 2283 - 2288.
- Simeone, O., Bar-Ness, Y. & Spagnolini, U. (2004). Pilot-based channel estimation for ofdm systems by tracking the delay-subspace, *IEEE Transactions on Wireless Communications* Vol. 3(No. 1): 315-325.
- Song, B., Gui, L., Guan, Y. & Zhang, W. (2005). On channel estimation and equalization in tds-ofdm based terrestrial hdtv broadcasting system, *IEEE Transactions Consumer Electronics* Vol. 51(No. 3): 790-797.
- Sourour, E. A. & Nakagawa, M. (1996). Performance of orthogonal multicarrier cdma in a multipath fading channel, *IEEE Transactions on Communications* Vol. 44(No. 3): 356 - 367.
- Stark, H. & Woods, J. W. (2001). *Probability and Random Processes with Applications to Signal Processing*, Prentice Hall, 3rd ed.
- Steele, R. (1999). *Mobile Radio Communications*, John Wiley & Sons, Inc.
- Tourtier, P. J., Monnier, R. & Lopez, P. (1993). Multicarrier model for digital hdtv terrestrial broadcasting, *Signal processing: Image communication* Vol. 5(No. 5-6): 379-403.
- Tu, J. C. (1991). *Theory, Design and Application of Multi-Channel Modulation for Digital Communications*, Ph. D. Dissertation, Stanford University, CA.
- Tufvesson, F. & Maseng, T. (1997). Pilot assisted channel estimation for ofdm in mobile cellular systems, *Proceedings of 1997 IEEE 47th Vehicular Technology Conference*, IEEE Vehicular Technology Society, Phoenix, AZ, pp. 1639-1643.
- Van de Beek, J.-J., Edfors, O., Sandell, M., Wilson, S. K. & Borjesson, P. O. (1995). On channel estimation in ofdm systems, *Proceedings of 1995 IEEE 45th Vehicular Technology Conference*, IEEE Vehicular Technology Society, Chicago, IL, pp. 815-819.
- Weinstein, S. B. & Ebert, P. M. (1971). Data transmission by frequency-division multiplexing using the discrete fourier transform, *IEEE Transactions on Communication Technology* Vol. 19(No. 5): 628-634.
- Wilson, S. K., Khayata, R. E. & Cioffi, J. M. (1994). 16-qam modulation with orthogonal frequency-division multiplexing in a rayleigh-fading environment, *Proceedings of*

- 1994 *IEEE 44th Vehicular Technology Conference*, IEEE Vehicular Technology Society, Stockholm, pp. 1660-1664.
- Yang, F., Wang, J., Wang, J., Song, J. & Yang, Z. (2008). Novel channel estimation method based on pn sequence reconstruction for chinese dttb system, *IEEE Transactions Consumer Electronics* Vol. 54(No. 4): 1583-1589.
- Yeh, C.-S. & Lin, Y. (1999). Channel estimation techniques based on pilot arrangement in ofdm systems, *IEEE Transactions on Broadcasting* Vol. 45(No. 4): 400-409.
- Young, G., Foster, K. T. & Cook, J. W. (1996). Broadband multimedia delivery over copper, *Electronics & Communication Engineering Journal* Vol. 8(No. 1): 25.
- Zhao, Y. & Huang, A. (1997). A novel channel estimation method for ofdm mobile communication systems based on pilot signals and transform-domain processing, *Proceedings of 1997 IEEE 47th Vehicular Technology Conference*, IEEE Vehicular Technology Society, Phoenix, AZ, pp. 2089-2093.
- Zheng, Z.-W. & Sun, Z.-G. (2008). Robust channel estimation scheme for the tds-ofdm based digital television terrestrial broadcasting system, *IEEE Transactions Consumer Electronics* Vol. 54(No. 4): 1576-1582.
- Zou, W. Y. & Wu, Y. (1995). Cofdm: an overview, *IEEE Transactions on Broadcasting* Vol. 41(No. 1): 1-8.

OFDM Communications with Cooperative Relays

H. Lu¹, H. Nikoogar¹ and T. Xu²

¹*International Research Centre for Telecommunications and Radar (IRCTR)*

²*Circuits and Systems Group (CAS)*

*Dept. EEMCS, Delft University of Technology
Mekelweg 4, 2628 CD, Delft,
The Netherlands*

1. Introduction

1.1 Cooperative relay communications

Signal fading due to multi-path propagation is one of the major impairments to meet the demands of next generation wireless networks for high data rate services. To mitigate the fading effects, time, frequency, and spatial diversity techniques or their hybrid can be used. Among different types of diversity techniques, spatial diversity is of special interest as it does not incur system losses in terms of delay and bandwidth efficiency.

Recently, cooperative diversity in wireless network has received great interest and is regarded as a promising technique to mitigate multi-path fading, which results in a fluctuation in the amplitude of the received signal. Cooperative communications is a new communication paradigm which generates independent paths between the user and the base station by introducing a relay channel. The relay channel can be thought of as an auxiliary channel to the direct channel between the source and destination. The basic idea behind cooperation is that several users in a network pool their resources in order to form a virtual antenna array which creates spatial diversity (Laneman et al., 2004; Sendonaris et al., Part I, 2003; Sendonaris et al., Part II, 2003). Since the relay node is usually several wavelengths distant from the source, the relay channel is guaranteed to fade independently from the direct channel, which introduces a full-rank Multiple-input-multiple-output (MIMO) channel between the source and the destination. This cooperative spatial diversity leads to an increased exponential decay rate in the error probability with increasing signal-to-noise ratio (SNR) (Liu et al., 2009).

Before discussing cooperative OFDM, let us first review some fundamental knowledge of OFDM and MIMO, which is associated with the cooperative OFDM study in this chapter.

1.2 Physical layer of cooperative wireless networks (OFDM & MIMO)

1.2.1 OFDM basics

In the modern wireless communication, OFDM technology has been widely used due to its spectral efficiency and inherent flexibility in allocating power and bit rate over distinct subcarriers which are orthogonal to each other. Different from a serial transmission, OFDM

is a multi-carrier block transmission, where, as the name suggests, information-bearing symbols are processed in blocks at both the transmitter and the receiver.

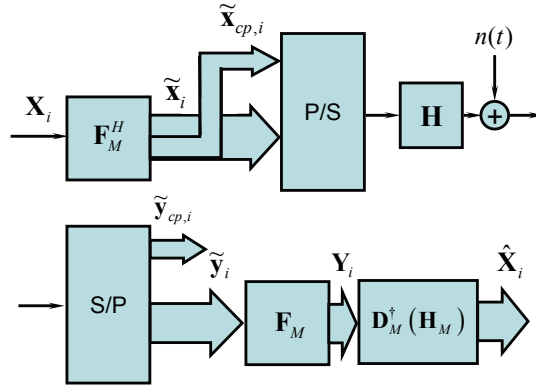


Fig. 1. Discrete-time block equivalent models of CP-OFDM, top: transmitter & channel, bottom: receiver.

A number of benefits the OFDM brings to cooperative relay systems originate from the basic features that OFDM possesses. To appreciate those, we first outline Cyclic Prefix (CP)-OFDM's operation using the discrete-time baseband equivalent block model of a single-transceiver system depicted in Fig.1, where \mathbf{X}_i is the so-called frequency signal at the i -th time symbol duration in one OFDM frame, then it will be transferred as $\tilde{\mathbf{x}}_i$ in the time domain by the M -point inverse fast Fourier transform (IFFT) matrix $\mathbf{F}_M^{-1} = \mathbf{F}_M^H$ with (m, k) -th entry $\exp(j2\pi mk / M) / \sqrt{M}$, i.e., $\tilde{\mathbf{x}}_i = \mathbf{F}_M^H \mathbf{X}_i$, \mathbf{F}_M is the M -point fast Fourier transform (FFT) matrix. where $(\cdot)^H$ denotes conjugate transposition, $(\cdot)^\dagger$ denotes matrix pseudoinverse, and $(\cdot)^{-1}$ denotes matrix inversion and m, k denote the index in frequency and time domain, respectively. Applying the triangle inequality to the M -point IFFT definition shows that the entries of $\mathbf{F}_M^H \mathbf{X}_i$ have magnitudes that can exceed those of \mathbf{X}_i by a factor as high as M . In other words, IFFT processing can increase the peak to average power ratio (PAPR) by a factor as high as the number of subcarriers (which in certain applications can exceed 1000). Then a CP of length D is inserted between each $\tilde{\mathbf{x}}_i$ to form the redundant OFDM symbols $\tilde{\mathbf{x}}_{cp,i}$, which are sequentially transmitted through the channel. The total number of the time domain signals in each OFDM symbol is, thus, $C = M + D$. If we define $\mathbf{F}_{cp} := [\mathbf{F}_D, \mathbf{F}_M]^H$ as the $C \times M$ expanded IFFT matrix, where \mathbf{F}_D is the last D columns of \mathbf{F}_M that way, the redundant OFDM symbol to be transmitted can also be expressed as $\tilde{\mathbf{x}}_{cp,i} = \mathbf{F}_{cp} \mathbf{X}_i$. With $(\cdot)^T$ denotes transposition, and assuming no channel state information (CSI) to be available at the transmitter, then the received symbol $\tilde{\mathbf{y}}_{cp,i}$ at the i -th time symbol duration can be written as:

$$\tilde{\mathbf{y}}_{cp,i} = \mathbf{H} \mathbf{F}_{cp} \mathbf{X}_i + \mathbf{H}_{ISI} \mathbf{F}_{cp} \mathbf{X}_{i-1} + \tilde{\mathbf{n}}_{C,i} \quad (1)$$

where \mathbf{H} is the $C \times C$ lower triangular Toeplitz filtering matrix with first column $[h_1 \cdots h_L \ 0 \cdots 0]^T$, where L is the channel order (i.e., $h_i = 0, \forall i > L$), \mathbf{H}_{ISI} is the $C \times C$ upper triangular Toeplitz filtering matrix with first row $[0 \cdots 0 \ h_L \cdots h_2]$, which captures inter-

symbol interference (ISI), $\tilde{\mathbf{n}}_{C,i}$ denotes the additive white Gaussian noise (AWGN) vector with variance N_0 and Length C . After removing the CP at the receiver, ISI is also discarded, and (1) can be rewritten as:

$$\tilde{\mathbf{y}}_i = \mathbf{C}_M(\mathbf{h})\mathbf{F}_M^H\mathbf{X}_i + \tilde{\mathbf{n}}_{M,i} \quad (2)$$

where $\mathbf{C}_M(\mathbf{h})$ is $M \times M$ circulant matrix with first row $[h_1 \ 0 \cdots 0 \ h_L \cdots h_2]$, and $\tilde{\mathbf{n}}_{M,i}$ is a vector formed by the last M elements of $\tilde{\mathbf{n}}_{C,i}$.

The procedure of adding and removing CP forces the linear convolution with the channel impulse response to resemble a circular convolution. Equalization of CP-OFDM transmissions ties to the well known property that a circular convolution in the time domain, is equivalent to a multiplication operation in the frequency domain. Hence, the circulant matrix can be diagonalized by post- (pre-) multiplication by (I)FFT matrices, and only a single-tap frequency domain equalizer is sufficient to resolve the multipath effect on the transmitted signal. After demodulation with the FFT matrix, the received signal is given by:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{F}_M\mathbf{C}_M(\mathbf{h})\mathbf{F}_M^H\mathbf{X}_i + \mathbf{F}_M\tilde{\mathbf{n}}_{M,i} \\ &= \text{diag}(H_1 \cdots H_M)\mathbf{X}_i + \mathbf{F}_M\tilde{\mathbf{n}}_{M,i} \\ &= \mathbf{D}_M(\mathbf{H}_M)\mathbf{X}_i + \mathbf{n}_{M,i} \end{aligned} \quad (3)$$

where $\mathbf{H}_M = [H_1 \cdots H_M]^T = \sqrt{M}\mathbf{F}_M\mathbf{h}$, with

$$H_k \equiv H(2\pi k / M) := \sum_{l=1}^L h_l e^{-j2\pi kl / M} \quad (4)$$

denoting the channel's transfer function on the k -th subcarrier, $\mathbf{D}_M(\mathbf{H}_M)$ stands for the $M \times M$ diagonal matrix with \mathbf{H}_M on its diagonal, $\mathbf{n}_{M,i} := \mathbf{F}_M\tilde{\mathbf{n}}_{M,i}$.

Equations (3) and (4) show that an OFDM system which relies on M subcarriers to transmit the symbols of each block \mathbf{X}_i , converts an FIR frequency-selective channel to an equivalent set of M flat fading subchannels. This is intuitively reasonable since each narrowband subcarrier that is used to convey each information-bearing symbol per OFDM block "sees" a narrow portion of the broadband frequency-selective channel which can be considered frequency flat. This scalar model enables simple equalization of the FIR channel (by dividing (3) with the corresponding scalar subchannel \mathbf{H}_M) as well as low-complexity decoding across subchannels using (Muquet et al., 2009; Wang & Giannakis, 2000). Transmission of symbols over subcarriers also allows for a flexible allocation of the available bandwidth to multiple users operating with possibly different rate requirements imposed by multimedia applications, which may include communication of data, audio, or video. When CSI is available at the transmitter side, power and bits can be adaptively loaded per OFDM subcarrier, depending on the strength of the intended subchannel. Because of orthogonality of OFDM subcarriers, OFDM system exhibits robustness to the narrow band interference.

The price paid for OFDM's attractive features in equalization, decoding, and possibly adaptive power and bandwidth allocation is its sensitivity to subcarrier drifts and the high PAPR that IFFT processing introduces to the entries of each block transmitted. Subcarrier

drifts come either from the carrier-frequency and phase offsets between transmit-receive oscillators or from mobility-induced Doppler effects, with the latter causing a spectrum of frequency drifts. Subcarrier drifts cause inter-carrier interference (ICI), which renders (3) invalid. On the other hand, high PAPR necessitates backing-off transmit-power amplifiers to avoid nonlinear distortion effects (Batra et al., 2004).

However, the same multipath robustness can be obtained by adopting ZP instead of CP (Lu et al., 2009). If the length of the zero-padding equals the length of CP, then the ZP-OFDM will achieve the same spectrum efficiency as CP-OFDM.

The only difference between the transmission part of the ZP-OFDM and CP-OFDM, as shown in Fig. 2, is the CP replaced by D appending zeros at the end of the symbol. If we define $\mathbf{F}_{zp} := [\mathbf{F}_M, \mathbf{0}]^H$, and $Z = C = M + D$, the transmitted OFDM symbol can be denoted as $\tilde{\mathbf{x}}_{zp,i} = \mathbf{F}_{zp} \mathbf{X}_i$. The received symbol is now expressed as:

$$\tilde{\mathbf{y}}_{zp,i} = \mathbf{H} \mathbf{F}_{zp} \mathbf{X}_i + \mathbf{H}_{ISL} \mathbf{F}_{zp} \mathbf{X}_{i-1} + \tilde{\mathbf{n}}_{z,i}. \quad (5)$$

The key advantage of ZP-OFDM relies on two aspects: first, the all-zero $D \times M$ matrix $\mathbf{0}$ is able to take good care of the ISI, when the length of the padded zeros is not less than the maximum channel delay. Second, according to the Eq. (4), multipath channel will introduce 3 impact factors, h_l , k and l to the received signal, which stand for the amplitude, subcarriers (in frequency domain) and delay (in time domain), respectively. Therefore, different CP copies from multipath certainly pose stronger interference than ZP copies. Thus, without equalization or some pre-modulation schemes, like Differential-PSK, the ZP-OFDM has a natural better bit error rate (BER) performance than the CP-OFDM. Furthermore, the linear structure of the channel matrix in ZP-OFDM ensures the symbol recovery regardless of the channel zeros locations.

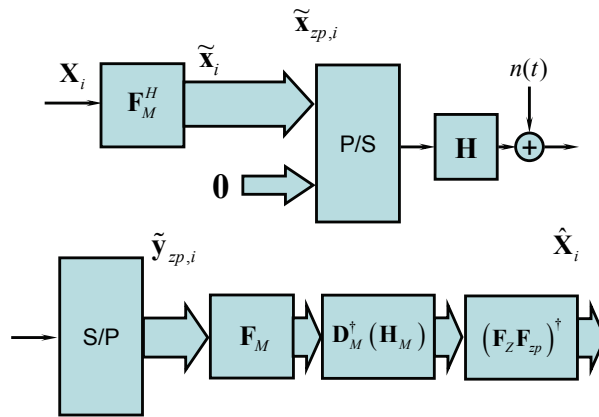


Fig. 2. Discrete-time block equivalent models of ZP-OFDM, top: transmitter & channel, bottom: receiver.

Nevertheless, because of the zero-padding and linear structure of ZP-OFDM, it outperforms CP-OFDM in terms of the lower PAPR (Batra et al., 2004; Lu et al., 2009). Similar to silent periods in TDMA, trailing zeros will not pose problems to high-power amplifiers (HPA). By adopting the proper filter, they will not give rise to out-of-band spectral leakage, either. The

circulant channel convolution matrix $\mathbf{C}_M(\mathbf{h})$ in the CP-OFDM is invertible if and only if the channel transfer function has no zeros on the FFT grid, i.e., $H_k \neq 0, \forall k \in [1, M]$, therefore, when channel nulls hit the transmitted symbols, the signal recovery becomes impossible. However, in the ZP-OFDM, the tall Toeplitz structure of equivalent channel matrix always guarantees its full rank (it only becomes rank deficient when the channel impulse response is identically zero, which is impossible in practice) (Muquet et al., 2009). In other words, the full rank property guarantees the detection of transmitted symbols.

In the blind channel estimation and blind symbol synchronization, ZP-OFDM also has its advantage in reducing the system complexity. Therefore, for more efficient utilization of the spectrum and low power transmission, a fast-equalized ZP-OFDM seems more promising than the CP-OFDM.

The above reviewed advantages and limitations of single-transceiver CP-OFDM and ZP-OFDM systems are basically present in the cooperative scenario which we present later under the name of cooperative OFDM.

1.2.2 From MIMO to cooperative communications

MIMO systems have been constructed comprising multiple antennas at both the transmitter and receiver to offer significant increases in data throughput and link range without additional expenditure in frequency and time domain. The spatial diversity has been studied intensively in the context of MIMO systems (Barbarossa, 2005). It has been shown that utilizing MIMO systems can significantly improve the system throughput and reliability (Foschini & Gans, 1998).

In the fourth generation wireless networks to be deployed in the next couple of years, namely, mobile broadband wireless access (MBWA) or IEEE 802.20, peak data rates of 260 Mbps can be achieved on the downlink, and 60 Mbps on the uplink (Hwang et al., 2007). These data rates can, however, only be achieved for full-rank MIMO users. More specifically, full-rank MIMO users must have multiple antennas at the mobile terminal, and these antennas must see independent channel fades to the multiple antennas located at the base station. In practice, not all users can guarantee such high rates because they either do not have multiple antennas installed on their small-size devices, or the propagation environment cannot support MIMO because, for example, there is not enough scattering. In the latter case, even if the user has multiple antennas installed full-rank MIMO is not achieved because the paths between several antenna elements are highly correlated.

To overcome the above limitations of achieving MIMO gains in future wireless networks, we must think of new techniques beyond traditional point-to-point communications. The traditional view of a wireless system is that it is a set of nodes trying to communicate with each other. From another point of view, however, because of the broadcast nature of the wireless channel, we can think of those nodes as a set of antennas distributed in the wireless system. Adopting this point of view, nodes in the network can cooperate together for a distributed transmission and processing of information. The cooperating node acts as a relay node for the source node. Since the relay node is usually several wavelengths distant from the source, the relay channels are guaranteed to fade independently from the direct channels, as well as each other which introduces a full-rank MIMO channel between the source and the destination. In the cooperative communications setup, there is a-priori few constraints to different nodes receiving useful energy that has been emitted by another transmitting node. The new paradigm in user cooperation is that, by implementing the appropriate signal

processing algorithms at the nodes, multiple terminals can process the transmissions overheard from other nodes and be made to collaborate by relaying information for each other. The relayed information is subsequently combined at a destination node so as to create spatial diversity. This creates a network that can be regarded as a system implementing a distributed multiple antenna where collaborating nodes create diverse signal paths for each other (Liu et al., 2009). Therefore, we study the cooperative relay communication system, and consequently, a cooperative ZP-OFDM to achieve the full diversity is investigated.

The rest of the chapter is organized as follows. In Section II, we first provide and discuss the basic models of AF, DF and their hybrid scheme. The performance analysis of the hybrid DF-AF is presented in Section III. The cooperative ZP-OFDM scheme, which will be very promising for the future cooperative Ultra Wide Band (UWB) system, is addressed in Section IV, the space time frequency coding (STFC) scheme for the full diversity cooperation is proposed as well. The conclusions of the chapter appear in Section VI.

2. System model

Cooperative communications is a new paradigm shift for the fourth generation wireless system that will guarantee high data rates to all users in the network, and we anticipate that it will be the key technology aspect in the fifth generation wireless networks (Liu et al., 2009).

In terms of research ascendance, cooperative communications can be seen as related to research on relay channel and MIMO systems. The concept of user cooperation itself was introduced in two-part series of papers (Sendonaris et al., Part I, 2003; Sendonaris et al., Part II, 2003). In these works, Sendonaris *et al.* proposed a two-user cooperation system, in which pairs of terminals in the wireless network are coupled to help each other forming a distributed two-antenna system. Cooperative communications allows different users or nodes in a wireless network to share resources and to create collaboration through distributed transmission/processing, in which each user's information is sent out not only by the user but also by the collaborating users (Nosratinia et al., 2004). Cooperative communications promises significant capacity and multiplexing gain increase in the wireless system (Kramer et al., 2005). It also realizes a new form of space diversity to combat the detrimental effects of severe fading. There are mainly two relaying protocols: AF and DF.

2.1 Amplify and forward protocol

In AF, the received signal is amplified and retransmitted to the destination. The advantage of this protocol is its simplicity and low cost implementation. But the noise is also amplified at the relay. The AF relay channel can be modeled as follows. The signal transmitted from the source x is received at both the relay and destination as

$$y_{s,r} = \sqrt{E_s} h_{s,r} x + n_{s,r}, \text{ and } y_{s,d} = \sqrt{E_s} h_{s,d} x + n_{s,d} \quad (6)$$

where $h_{s,r}$ and $h_{s,d}$ are the channel gains between the source and the relay and destination, respectively, and are modeled as Rayleigh flat fading channels. The terms $n_{s,r}$ and $n_{s,d}$ denote the additive white Gaussian noise with zero-mean and variance N_0 , E_s is the average transmission energy at the source node. In this protocol, the relay amplifies the signal from the source and forwards it to the destination ideally to equalize the effect of the channel

fading between the source and the relay. The relay does that by simply scaling the received signal by a factor A_r that is inversely proportional to the received power, which is denoted by

$$A_r = \sqrt{\frac{E_s}{E_s h_{s,r} + N_0}} \quad (7)$$

The destination receives two copies from the signal x through the source link and relay link. There are different techniques to combine the two signals at the destination. The optimal technique that maximizes the overall SNR is the maximal ratio combiner (MRC). Note that the MRC combining requires a coherent detector that has knowledge of all channel coefficients, and the SNR at the output of the MRC is equal to the sum of the received signal-to-noise ratios from all branches.

2.2 Decode and forward protocol

Another protocol is termed as a decode-and-forward scheme, which is often simply called a DF protocol. In the DF, the relay attempts to decode the received signals. If successful, it re-encodes the information and retransmits it. Although DF protocol has the advantage over AF protocol in reducing the effects of channel interferences and additive noise at the relay, the system complexity will be increased to guarantee the correct signal detection.

Note that the decoded signal at the relay may be incorrect. If an incorrect signal is forwarded to the destination, the decoding at the destination is meaningless. It is clear that for such a scheme the diversity achieved is only one, because the performance of the system is limited by the worst link from the source-relay and source-destination (Laneman et al., 2004).

Although DF relaying has the advantage over AF relaying in reducing the effects of noise and interference at the relay, it entails the possibility of forwarding erroneously detected signals to the destination, causing error propagation that can diminish the performance of the system. The mutual information between the source and the destination is limited by the mutual information of the weakest link between the source-relay and the combined channel from the source-destination and relay-destination.

Since the reliable decoding is not always available, which also means DF protocol is not always suitable for all relaying situations. The tradeoff between the time-consuming decoding, and a better cooperative transmission, finding the optimum hybrid cooperative schemes, that include both DF and AF for different situations, is an important issue for the cooperative wireless networks design.

2.3 Hybrid DF-AF protocol

In this chapter, we consider a hybrid cooperative OFDM strategy as shown in Fig. 3, where we transmit data from source node S to destination node D through R relays, without the direct link between S and D . This relay structure is called 2-hop relay system, i.e., first hop from source node to relay, and second hop from relay to destination. The channel fading for different links are assumed to be identical and statistically independent, quasi-statistic, i.e., channels are constant within several OFDM symbol durations. This is a reasonable assumption as the relays are usually spatially well separated and in a slow changing environment. We assume that the channels are well known at the corresponding receiver

sides, and a one bit feedback channel from destination to relay is used for removing the unsuitable AF relays. All the AWGN terms have equal variance N_0 . Relays are re-ordered according to the descending order of the SNR between S and Q , i.e., $\text{SNR}_{sQ_1} > \dots > \text{SNR}_{sQ_R}$, where SNR_{sQ_r} denotes the r -th largest SNR between S and Q .

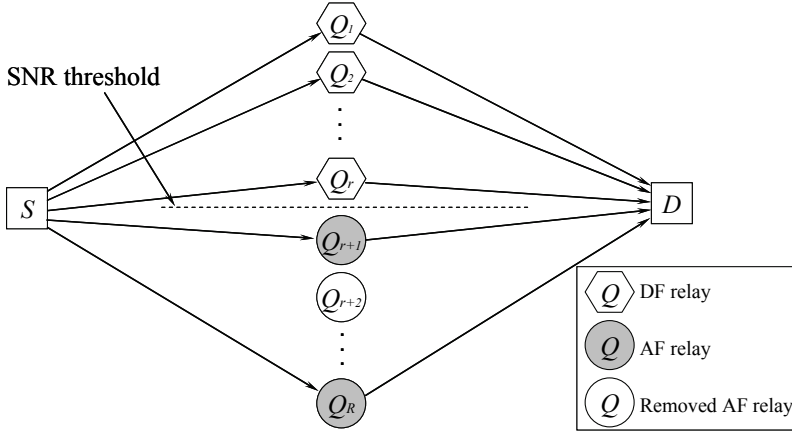


Fig. 3. Hybrid relay cooperation with dynamic optimal combination of DF-AF relays (S : Source, D : Destination, Q_r : r -th Relay)

In this model, relays can determine whether the received signals are decoded correctly or not, just simply by comparing the SNR to the threshold, which will be elaborated in Section 3.1. Therefore, the relays with SNR above the threshold will be chosen to decode and forward the data to the destination, as shown with the white hexagons in Fig.3. The white circle is the removed AF relay according to the dynamic optimal combination strategy which will be proposed in Section 3.2. The rest of the relays follow the AF protocol, as shown with the white hexagons in Fig. 3 (Lu & Nikookar, 2009; Lu et al., 2010).

The received SNR at the destination in the hybrid cooperative network can be denoted as

$$\gamma_h = \sum_{Q_i \in \text{DF}} \frac{E_Q h_{Q_i, D}}{N_0} + \sum_{Q_j \in \text{AF}} \frac{\frac{E_s h_{s, Q_j}}{N_0} \frac{E_Q h_{Q_j, D}}{N_0}}{\frac{E_s h_{s, Q_j}}{N_0} + \frac{E_Q h_{Q_j, D}}{N_0} + 1} \quad (8)$$

where $h_{Q_i, D}$, h_{s, Q_j} and $h_{Q_j, D}$ denote the power gains of the channel from the i -th relay to the destination in DF protocol, source node to the j -th relay in AF protocol and j -th relay to the destination in AF protocol, respectively. E_s and E_Q in (8) are the average transmission energy at the source node and at the relays, respectively. By choosing the amplification factor A_{Q_j} in the AF protocol as:

$$A_{Q_j}^2 = \frac{E_s}{E_s h_{s, Q_j} + N_0} \quad (9)$$

and forcing the E_Q in DF equal to E_S , it will be convenient to maintain constant average transmission energy at relays, equal to the original transmitted energy at the source node. In this chapter, OFDM is used as a modulation technique in the cooperative system to gain from its inherent advantages and combat frequency selective fading of each cooperative link, with W_r , $r = 1, 2, \dots, R$ independent paths. Later, we also show that, by utilizing the space-frequency coding, hybrid DF-AF cooperative OFDM can also gain from the frequency selective fading and achieve the multi-path diversity with a diversity gain of $W_{min} = \min(W_r)$. As shown in the Fig.4, the r -th relay first decides to adopt DF or AF protocol according to the SNR threshold. For the DF-protocol, the symbols are decoded at the relays, and then an IFFT operation is applied on these blocks to produce the OFDM symbol. Before transmission, a prefix (CP or ZP) is added to each OFDM symbol. For the AF-protocol, relays which undergo the deep fading will be removed by using the dynamic optimal combination strategy discussed later in this section. Other AF relays are proper relays, amplify and forward the data to the destination. At the destination node, after the prefix removal, the received OFDM symbols are fast-Fourier-transformed, and the resulting symbols at the destination are used for the combination and detection.

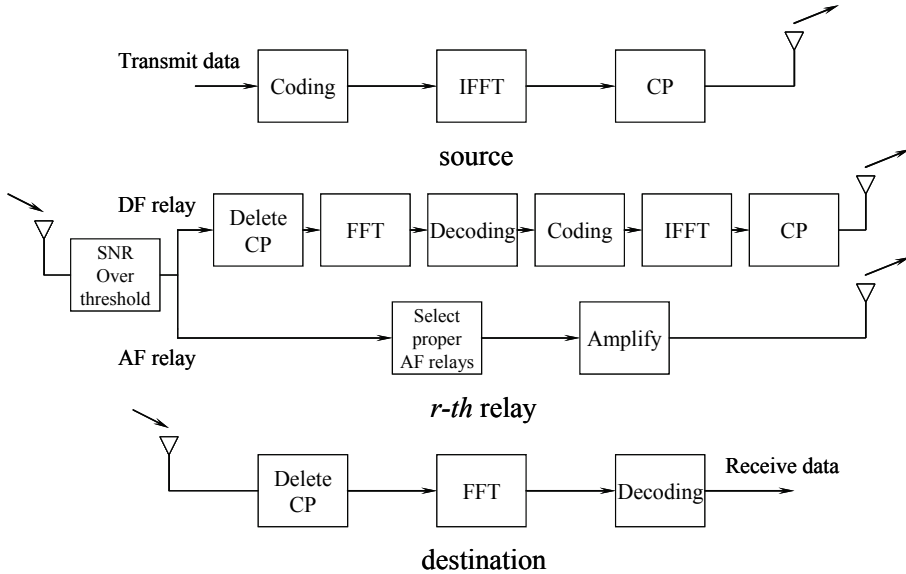


Fig. 4. Relay selection in the hybrid DF-AF cooperative OFDM wireless transmission strategy (top: source, middle: relay, bottom: destination)

The receiver at the destination collects the data from DF and AF relays with a MRC. Because of the amplification in the intermediate stage in the AF protocol, the overall channel gain of the AF protocol should include the source to relay, relay to destination channels gains and amplification factor. The decision variable u at the MRC output is given by

$$u = \sum_{Q_i \in \text{DF}} \frac{(H_{Q_i,D})^* Y_{Q_i}}{(H_{Q_i,D})^* H_{Q_i,D}} + \sum_{Q_j \in \text{AF}} \frac{(H_{S,Q_j} A_{Q_j} H_{Q_j,D})^* Y_{Q_j}}{(H_{S,Q_j} A_{Q_j} H_{Q_j,D})^* (H_{S,Q_j} A_{Q_j} H_{Q_j,D})} \quad (10)$$

where Y_{Q_i} and Y_{Q_j} are the received signal from DF i -th relay and AF j -th relays, respectively, and $(\cdot)^*$ denotes the conjugate operation. $H_{Q_i,D}$, H_{S,Q_j} and $H_{Q_j,D}$ are frequency response of the channel power gains, respectively.

In the proposed hybrid DF-AF cooperative network, DF plays a dominant role in the whole system. However, switching to AF scheme for the relay nodes with SNR below the threshold often improves the total transmission performance, and accordingly AF plays a positive compensating role.

3. Performance analysis of Hybrid DF-AF protocol

3.1 Threshold for DF and AF relays

In general, mutual information I is the upper bound of the target rate B bit/s/Hz, i.e., the spectral efficiency attempted by the transmitting terminal. Normally, $B \leq I$, and the case $B > I$ is known as the outage event. Meanwhile, channel capacity, C , is also regarded as the maximum achievable spectral efficiency, i.e., $B \leq C$.

Conventionally, the maximum average mutual information of the direct transmission between source and destination, i.e., I_D , achieved by independent and identically distributed (i.i.d) zero-mean, circularly symmetric complex Gaussian inputs, is given by

$$I_D = \log_2(1 + \text{SNR } h_{S,D}) \quad (11)$$

as a function of the power gain over source and destination, $h_{S,D}$. According to the inequality $B \leq I$, we can derive the SNR threshold for the full decoding as

$$\text{SNR} \geq \frac{2^B - 1}{h_{S,D}} \quad (12)$$

Then, we suppose all of the X relays adopt the DF cooperative transmission without direct transmission. The maximum average mutual information for DF cooperation I_{DF_co} is shown (Laneman et al., 2004) to be

$$I_{DF_co} = \frac{1}{X} \min \left\{ \log_2 \left(1 + \sum_{r=1}^R \text{SNR } h_{S,Q_r} \right), \log_2 \left(1 + \sum_{r=1}^R \text{SNR } h_{Q_r,D} \right) \right\} \quad (13)$$

which is a function of the channel power gains. Here, R denotes the number of the relays. For the r -th DF link, requiring both the relay and destination to decode perfectly, the maximum average mutual information I_{DF_li} can be shown as

$$I_{DF_li} = \min \left\{ \log_2 \left(1 + \text{SNR } h_{S,Q_r} \right), \log_2 \left(1 + \text{SNR } h_{Q_r,D} \right) \right\} \quad (14)$$

The first term in (14) represents the maximum rate at which the relay can reliably decode the source message, while the second term in (14) represents the maximum rate at which the destination can reliably decode the message forwarded from relay. We note that such mutual information forms are typical of relay channel with full decoding at the relay (Cover & El Gamal, 1979). The SNR threshold of this DF link for target rate B is given by $I_{DF_li} \geq B$ which is derived as

$$\text{SNR} \geq \frac{2^B - 1}{\min(h_{S,Q_r}, h_{Q_r,D})} \quad (15)$$

In the proposed hybrid DF-AF cooperative transmission, we only consider that a relay can fully decode the signal transmitted over the source-relay link, but not the whole DF link. Thus, the SNR threshold for the full decoding at the r -th relay reaches its lower bound as

$$\gamma_{th} \geq \frac{2^B - 1}{h_{S,Q_r}} \quad (16)$$

For the DF protocol, let R denote the number of the total relays, M denote the set of participating relays, whose SNR_S are above the SNR threshold, and the reliable decoding is available. The achievable channel capacity, C_{DF} , with SNR threshold is calculated as

$$C_{DF} = \sum_M \frac{1}{R} \mathbb{E}(\log_2(1 + y|M)) \Pr(M) \quad (17)$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator, $y|M = (R-K)\gamma_{S,D} + \sum_{Q \in M} \gamma_{Q,D}$ denotes the instantaneous received SNR at the destination given set M with K participating relays, where $\gamma_{n,m}$ denotes the instantaneous received SNR at node m , which is directly transmitted from n to m . Since $y|M$ is the weighted sum of independent exponential random variables (Farhadi & Beaulieu, 2008), the probability density function (PDF) of $y|M$ can be obtained using its moment generating function (MGF) and partial fraction technique for evaluation of the inverse Laplace transform, see Eq. (8d) and Eq. (8e) in (Farhadi & Beaulieu, 2008). $\Pr(M)$ in (17) is the probability of a particular set of participating relays which are obtained as

$$\Pr(M) = \prod_{Q \in M} \exp\left(-\frac{R\gamma_{th}}{\Gamma_{S,Q \in M}}\right) \prod_{Q \notin M} \left(1 - \exp\left(-\frac{R\gamma_{th}}{\Gamma_{S,Q \notin M}}\right)\right) \quad (18)$$

where $\Gamma_{u,v}$ denotes the average SNR over the link between nodes u and v .

Combining (13), (17) and (18) with the inequality $I_{DF-co} \leq C_{DF}$, since the maximum average mutual information, I , is upper bound by the achievable channel capacity, C , we can calculate the upper bound of SNR threshold γ_{th} for fully decoding in the DF protocol.

Now, we can obtain the upper bound and the lower bound of the SNR threshold γ_{th} for the hybrid DF-AF cooperation. However, compared to the upper bound, the lower bound as shown in the (16) is more crucial for improving the transmission performance. This is because the DF protocol plays a dominant role in the hybrid cooperation strategy, and accordingly we want to find the lower bound which provides as much as possible DF relays. We will elaborate this issue later. Fully decoding check can also be guaranteed by employing the error detection code, such as cyclic redundancy check. However, it will increase the system complexity (Lin & Costello, 1983).

3.2 Dynamic optimal combination scheme

In the maximum ratio combining the transmitted signal from R cooperative relays nodes, which underwent independent identically distributed Rayleigh fading, and forwarded to

the destination node are combined. In this case the SNR per bit per relay link γ_r has an exponential probability density function (PDF) with average SNR per bit $\bar{\gamma}$:

$$p_{\gamma_r}(\gamma_r) = \frac{1}{\bar{\gamma}} e^{-\gamma_r/\bar{\gamma}} \quad (19)$$

Since the fading on the R paths is identical and mutually statistically independent, the SNR per bit of the combined SNR γ_c will have a Chi-square distribution with $2R$ degrees of freedom. The PDF $p_{\gamma_c}(\gamma_c)$ is

$$p_{\gamma_c}(\gamma_c) = \frac{1}{(R-1)! \bar{\gamma}_c^R} \gamma_c^{R-1} e^{-\gamma_c/\bar{\gamma}_c} \quad (20)$$

where $\bar{\gamma}_c$ is the average SNR per channel, then by integrating the conditional error probability over $\bar{\gamma}_c$, the average probability of error P_e can be obtained as

$$P_e = \int_0^{\infty} \mathbb{Q}(\sqrt{2g\gamma_c}) p_{\gamma_c}(\gamma_c) d\gamma_c \quad (21)$$

where $g = 1$ for coherent BPSK, $g = 1/2$ for coherent orthogonal BFSK, $g = 0.715$ for coherent BFSK with minimum correlation, and $\mathbb{Q}(\cdot)$ is the Gaussian Q-function, i.e.,

$\mathbb{Q}(x) = 1/\sqrt{2\pi} \int_x^{\infty} \exp(-t^2/2) dt$. For the BPSK case, the average probability of error can be found in the closed form by successive integration by parts (Proakis, 2001), i.e.,

$$P_e = \left(\frac{1-\mu}{2}\right)^R \sum_{k=0}^{R-1} \binom{R-1+k}{k} \left(\frac{1+\mu}{2}\right)^k \quad (22)$$

where

$$\mu = \sqrt{\frac{\bar{\gamma}_c}{1+\bar{\gamma}_c}} \quad (23)$$

In the hybrid DF-AF cooperative network with two hops in each AF relay, the average SNR per channel $\bar{\gamma}_c$ can be derived as

$$\bar{\gamma}_c = \frac{\gamma_h}{K+2 \times J} \quad (24)$$

where K and J are the numbers of the DF relays and AF relays, respectively. γ_h can be obtained from (8). In the DF protocol, due to the reliable detection, we only need to consider the last hops, or the channels between the relay nodes and destination node.

As the average probability of error P_e is a precise indication for the transmission performance, we consequently propose a dynamic optimal combination strategy for the hybrid DF-AF cooperative transmission. In this algorithm the proper AF relays are selected to make P_e reach maximum.

First of all, like aforementioned procedure, relays are reordered according to the descending order of the SNR between source and relays, as shown in the Fig.3. According to the proposed SNR threshold, we pick up the DF relays having SNR greater than threshold. Then, we proceed with the AF relay selection scheme, where the inappropriate AF relays are removed. The whole dynamic optimal combination strategy for the hybrid DF-AF cooperation is shown in the flow chart of Fig. 5.

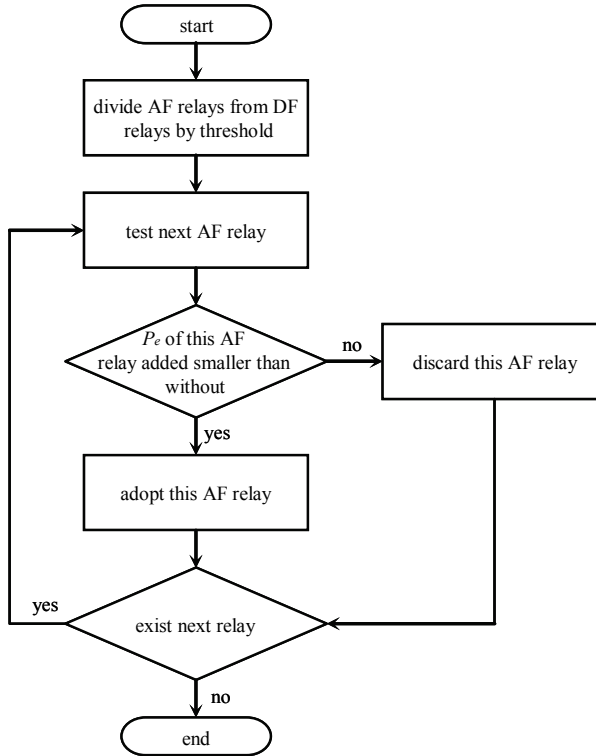


Fig. 5. Flow chart of the dynamic optimal combination strategy for the hybrid DF-AF cooperation

By exploiting the space-frequency coding proposed in (Li et al., 2009), we can further gain from the hybrid DF-AF cooperative OFDM in the frequency selective channel by coding across relays and OFDM tones, and obtain the multi-path diversity. According to the Eq. (14) in (Li et al., 2009), the multi-path diversity of the hybrid DF-AF cooperative OFDM can be shown by the upper bound of the error probability as:

$$\begin{aligned}
 P_e &< G_c^{-RW_{\min}} \left(\frac{\log \gamma_h}{\gamma_h} \right)^{RW_{\min}} \\
 &\approx (G_c \gamma_h)^{-RW_{\min}} \text{ as } \gamma_h \rightarrow \infty
 \end{aligned} \tag{25}$$

where G_c is a constant, which can be shown as Eq. (35) in (Li et al., 2009), γ_h is the average SNR at the destination in the hybrid cooperative network, and can be calculated by (8) in this chapter.

It can be seen from (25) that the achievable diversity gain is RW_{min} , i.e., the product of the cooperative (relay) diversity R and the multi-path diversity W_{min} . Here $W_{min} = \min(W_r)$, where W_r , $r = 1, 2, \dots, R$ is the number of independent paths in each relay-destination link.

3.3 Simulation results

First, we simulated BPSK modulation, Rayleigh channel, flat fading, without OFDM, and supposed the SNR threshold for correct decoding is $4E_b/N_0$, then we assumed $h_{Q_i,D} = h_{S,Q_j} = h_{Q_j,D} = 1$, for all branches, to verify proposed analytical BER expression. The resulting average BERs were plotted against the transmit SNR defined as $\text{SNR} = E_b/N_0$. As shown in the Fig. 6, the theoretical curves of multi-DF cooperation derived from our analytical closed-form BER expression clearly agree with the Monte Carlo simulated curves, while the theoretical curves of 2-AF and 3-AF cooperation match the simulation result only at the low SNR region.

Fig. 7 shows the BER performance for hybrid DF-AF cooperation. For the DF-dominant hybrid cooperation, the theoretical curves exhibit a good match with the Monte Carlo simulation results curves. The slight gap between theoretical and simulation BER results for the hybrid case of 1-DF + multi-AF can be explained by the AF relay fading which was considered as a double Gaussian channel, a product of two complex Gaussian channel (Patel et al., 2006). Obviously, the distribution of combined SNR (i.e., γ_c) will no longer follow the chi-square distribution giving rise to this slight difference.

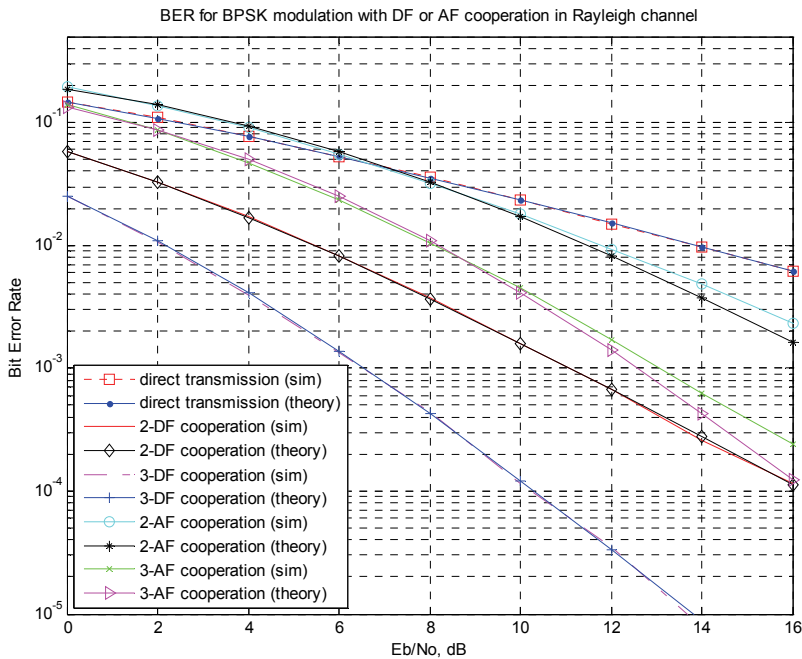


Fig. 6. BER performance for DF or AF cooperation.

In this proposed hybrid cooperation protocol, DF is dominant. We show this characteristic of the hybrid DF-AF cooperation by the following theorem:

Theorem 1: For the F -hop relay link, and the full decoding in DF protocol, as long as the SNR of the last hop is larger than $1/F$ times of the arithmetic mean of the whole link SNR, DF always plays a more important role than AF in improving the BER performance.

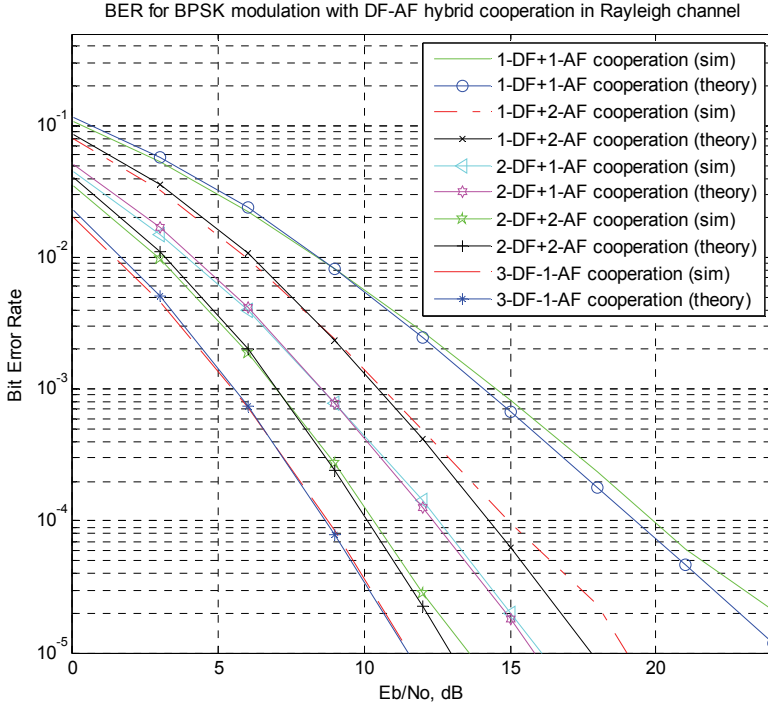


Fig. 7. BER performance for hybrid DF+AF cooperation.

Proof: According to the (22) and (23), the average probability of error P_e is a decreasing function w.r.t. combined SNR, γ_c . The SNR of the F -hop AF relay link, γ_{AF} , is the $1/F$ times of the harmonic mean of γ_i , $\forall i \in [1, F]$, i.e. (Hasna & Alouini, 2002),

$$\gamma_{AF} = \frac{\gamma_1 \gamma_2 \dots \gamma_L}{\sum_{i=1}^L \gamma_1 \gamma_2 \dots \gamma_{i-1} \gamma_{i+1} \dots \gamma_L} \quad (26)$$

Using Pythagorean means theorem, the harmonic mean is always smaller than the arithmetic mean. ■

For instance, in the high SNR region, the second term of (8) can be approximated as the $1/2$ times the harmonic mean of the 2-hop SNR in AF relay link (i.e., 1 is negligible in the denominator). As in practice, it is very easy for the last hop relay to achieve a SNR larger than $1/L$ times of the arithmetic mean of the whole link SNR, we can only consider the last hop of the reliably decoded DF protocol. Therefore, under the condition of the correct decoding, DF can enhance the error probability performance better than AF in the cooperative relay network.

This DF dominant hybrid cooperative networks strategy can be verified by the above simulation results as well. Comparing 2-DF to 2-AF in Fig. 6, or 2-DF plus 1-AF to 1-DF plus 2-AF in Fig. 7, or other hybrid DF-AF protocols with the same R , we can see that the fully decoded DF protocols always show a better BER performance than AF protocols. Therefore, DF protocols with a reliable decoding play a more important role in hybrid cooperative networks than AF protocols. Meanwhile, we can see from the figure that, changing to the AF scheme for the relay nodes with SNR below the threshold also improves the BER performance, as well as the diversity gain of the whole network. In fact, this is a better way than just discarding these relay nodes.

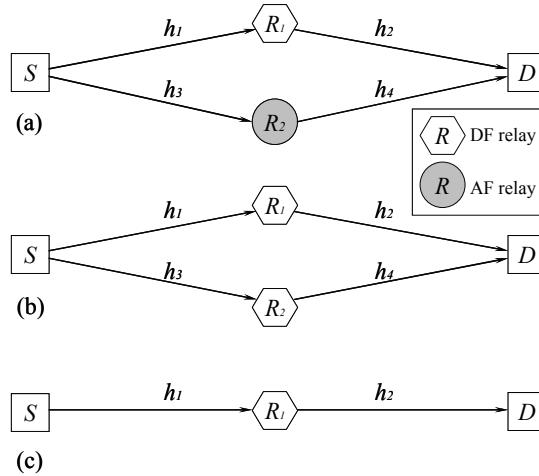


Fig. 8. hybrid DF-AF cooperation and DF cooperation architectures with different average power gains. (a) hybrid DF-AF cooperation, (b) dual DF cooperation, (c) single DF cooperation. (S: Source, D: Destination, h : average power gain between two nodes).

Reference (Louie et al., 2009) proposes a closed-form BER expression for two-hop AF protocol, which includes Gauss' hypergeometric and Gamma functions. This closed-form BER expression needs more computational burden to derive the cooperative analytical expression. In (Sadek et al., 2007), the analytical expression for multi-node DF protocol is provided with a complicated form as well. Instead, the compact closed-form BER expression for hybrid DF-AF cooperation proposed in this chapter allows us to achieve insight into the results with relatively low computations. The simple expressions can also help understanding the factors affecting the system performance. It can also be used for designing different network functions such as power allocation, scheduling, routing, and node selection.

In order to study the effect of the channel gains between source, relay and destination, we compare the hybrid DF-AF with the dual DF as well as the single DF cooperation in Fig. 8. In this figure, h_1 , h_2 , h_3 and h_4 stand for the average power gain between corresponding two nodes. In this simulation, the SNR threshold for correct decoding is assumed to be $4E_b/N_0$, and we set the first hop average power gain in DF protocol, i.e., h_1 in Fig. 8 (a) and Fig. 8 (c), and h_1 , h_3 in Fig. 8 (b) as 4, which means that the relay in DF protocol can fully decode the signal. The average power gains of the first hop in AF protocol, i.e., h_3 in Fig. 8 (a) increases

from 0.25 to 20. It can be seen from the Fig. 9 that the dual DF cooperation with reliable decoding outperforms the hybrid DF-AF cooperation, when corresponding average power gains are the same, i.e., diamond marked curve is better than square marked curve in Fig. 9. Meanwhile, the comparison of the curves shows that, the AF relay which undergoes the deep fading deteriorates the BER performance of hybrid DF-AF cooperation in the low SNR region. Thus, this AF relay should be removed according to the proposed dynamic optimal combination strategy to improve the transmission performance. Sum up the above discussion, due to power control, long transmission range, serious attenuation, etc., high SNR at relay and full decoding for DF protocol is not always available. In this case, relays can change to AF protocol with enough SNR to gain from the cooperative diversity.

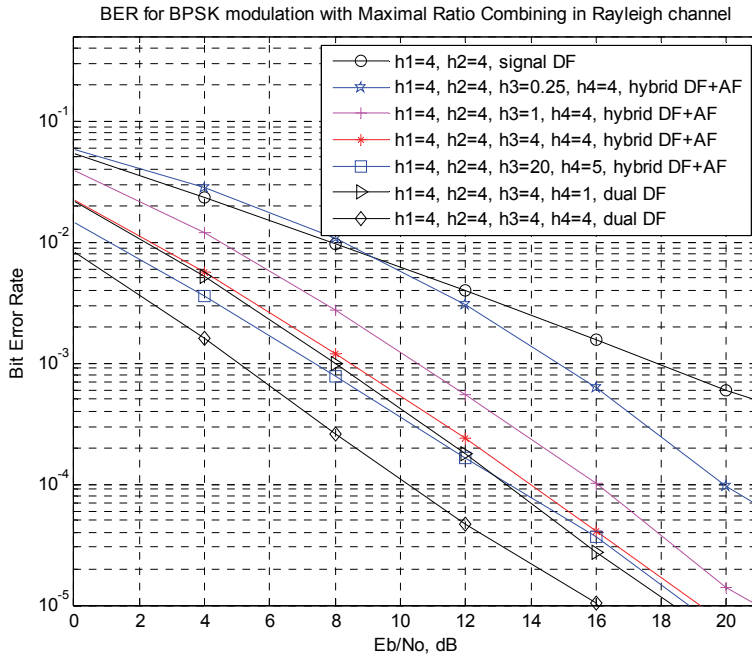


Fig. 9. BER performance for hybrid DF-AF cooperation and DF cooperation with different path gains.

Finally, we illustrate the validity of the theoretical results for the OFDM cooperation via simulations. An OFDM system with 64-point FFT and a CP length of 16 samples, which accounts for 25% of the OFDM symbol was considered. In the simulation, a more practical scenario was considered with a 3-path Rayleigh fading between each source node and relay node or relay node and destination node, i.e., $W_r = 3$. The 3-path delays were assumed at 0, 1, 2 samples, respectively. As illustrated in the Fig. 10, OFDM with CP can nicely cope with the multi-path, and the theoretical curves derived from (22) clearly agree with the Monte Carlo simulation curves. The simulation results indicate that under the condition of ISI resolved by OFDM and reliable decoding, the cooperative diversity gains from the increasing R , which is also shown by (8).

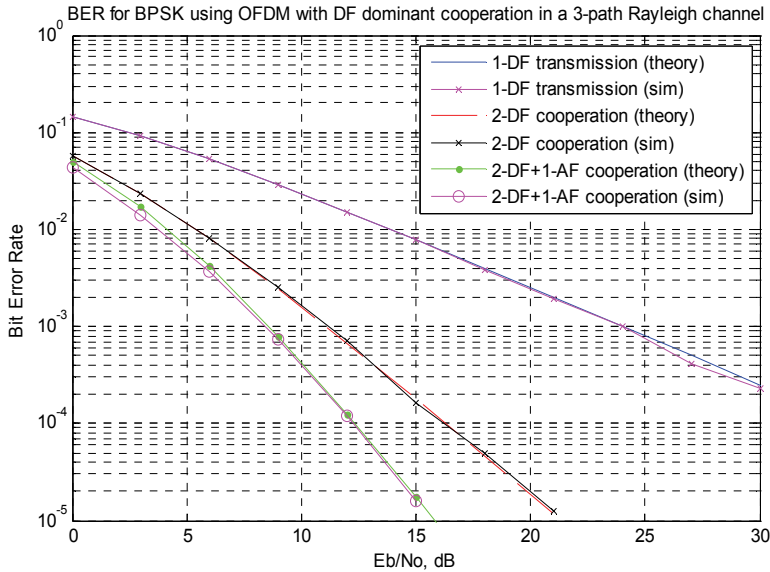


Fig. 10. BER performance for DF dominant OFDM cooperation.

4. ZP-OFDM with cooperative relays

4.1 CP and ZP for cooperative OFDM

Among the many possible multicarrier modulation techniques, OFDM is the one that has gained more acceptance as the modulation technique for high-speed wireless networks and 4G mobile broadband standards. Conventionally, CP is exploited to eliminate inter-symbol-interference (ISI) due to multipath. With a cyclic extension, the linear convolution channel is transformed into a circular convolution channel, and the ISI can be easily resolved. However, the cyclic prefix is not the only solution to combat the multipath. ZP has been recently proposed as an alternative to the CP in OFDM transmissions and Cognitive Radio (Lu et al., 2009). One of the advantages of using a ZP is that the transmitter requires less power back-off at the analog amplifier. Since the correlation caused by the cyclic prefix creates discrete spectral lines (ripples) into the average power spectral density of the transmitted signal and the radio emission power levels are limited by the Federal Communications Commission (FCC), the presence of any ripples in the power spectrum density (PSD) requires additional power back-off at the transmitter. In fact, the amount of power back-off that is required is equal to the peak-to-average ratio of the PSD.

A multiband (MB) ZP OFDM-based approach to design UWB transceivers has been recently proposed in (Batra et al., 2004) and (Batra, 2004) for the IEEE Standard. In Dec. 2008, the European Computer Manufacturers Association (ECMA) adopted ZP-OFDM for the latest version of High rate UWB Standard (Standard ECMA-368, 2008). Because of its advantage in the low power transmission, ZP-OFDM will have the potential to be used in other low power wireless communications systems.

We know that the multiple transmissions in the cooperative system may not be either time or frequency synchronized, i.e., signals transmitted from different transmitters arrive at the

receiver as different time instances, and multiple carrier frequency offsets (CFOs) also exist due to the oscillator mismatching. Different from the conventional MIMO system, the existence of multiple CFOs in the cooperative systems makes the direct CFOs compensation hard if not impossible. Therefore, in this section, we will investigate the cooperative ZP-OFDM system with multipath channel and CFOs, a subject that has not been addressed before. We propose a STFC, to hold the linear structure of the ZP-OFDM, and achieve the full cooperative spatial diversity, i.e., full multi-relay diversity. Furthermore, we show that, with only *linear receivers*, such as zero forcing (ZF) and minimum mean square error (MMSE) receivers, the proposed code achieves full diversity.

4.2 Full cooperative diversity with linear equalizer

4.2.1 Fundamental limits of diversity with linear equalizer

To quantify the performance of different communication systems, two important criteria are the average bit-error rates (BERs) and capacity. The BER performance of wireless transmissions over fading channels is usually quantified by two parameters: diversity order and coding gain (Liu et al., 2003; Tse & Viswanath, 2005). The diversity order is defined as the asymptotic slope of the BER versus signal-to-noise ratio (SNR) curve plotted in log-log scale. It describes how fast the error probability decays with SNR, while the coding gain measures the performance gap among different schemes when they have the same diversity. The higher the diversity, the smaller the error probability at high-SNR regimes. To cope with the deleterious effects of fading on the system performance, diversity-enriched transmitters and receivers have well-appreciated merits. Reference (Ma & Zhang, 2008) reveals the relationship between the channel orthogonality deficiency (*od*) and system full diversity. The orthogonality deficiency (*od*) indicates the degree of difficulty for the signal detection in the certain transceiver and channel condition (the smaller *od*, the easier signal detection). References (Shang & Xia, 2007; Shang & Xia, 2008) provide the two conditions for linear equalizer to achieve the full diversity. We will illustrate in this Section how to design the STFC to achieve full diversity for ZP-OFDM system.

In addition to focusing on diversity performance, practical systems also give high priority to reducing receiver complexity. Although maximum likelihood equalizer (MLE) enjoys the maximum diversity performance, its exponential decoding complexity makes it infeasible for certain practical systems. Some near-ML schemes (e.g., sphere decoding) can be used to reduce the decoding complexity. However, at low SNR or when large decoding blocks are sent/or high signal constellations are employed, the complexity of near-ML schemes is still high. In addition, these near-ML schemes adopt linear equalizers as preprocessing steps. To further reduce the complexity, when the system model is linear, one may apply linear equalizers (LEs) (Ma & Zhang, 2008).

4.2.2 System model and Linear ZP-OFDM

We consider a cooperative ZP-OFDM system as shown in the Fig. 11. Here the DF protocol is adopted in the cooperative communication model. Relays can fully decode the information, and participate in the cooperation, and occupy different frequency bands to transmit data to the destination. Each relay-destination link undergoes multipath Rayleigh fading. For the relay r , $r \in [1, 2, \dots, R]$, R is the number of relays, the received signal \mathbf{y}_r can be formulated as

$$\mathbf{y}_r = \mathbf{F}_{p,r} \mathbf{D}_{p,r} \mathbf{H}_r \mathbf{T}_{ZP} \mathbf{F}_{N,r}^H \mathbf{x}_r + \mathbf{n} \quad (27)$$

where $\mathbf{x}_r \triangleq [x_0, \dots, x_{N-1}]^T$ is the vector of the so called frequency transmitted information signal, and N is the signal length. The subscript r here indicates the variables or operators related to the r -th relay. To simplify the exposition, we only consider the effect of CFOs on signal. The noise term is denoted as \mathbf{n} , which stands for i.i.d. complex white Gaussian noise with zero mean. $\mathbf{F}_{N,r}$ stands for the N -point FFT matrix with (m, k) -th entry $\exp(j2\pi mk / N) / \sqrt{N}$, while $\mathbf{F}_{P,r}$ stands for the P -point FFT matrix.

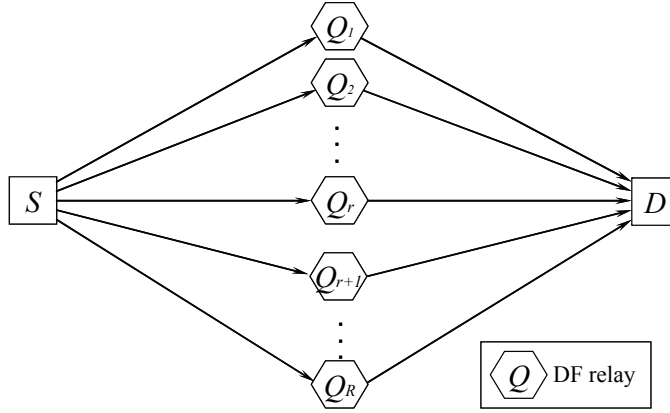


Fig. 11. Cooperative ZP-OFDM system architecture, (S : Source, D : Destination, Q_r : r -th Relay).

The matrix

$$\mathbf{T}_{ZP} = \begin{bmatrix} \mathbf{I}_N \\ 0 \end{bmatrix}_{P \times N} \quad (28)$$

performs the zero-padding on the transmitted signal with V zeros, where \mathbf{I}_N is $N \times N$ identity matrix, and $P = N + V$.

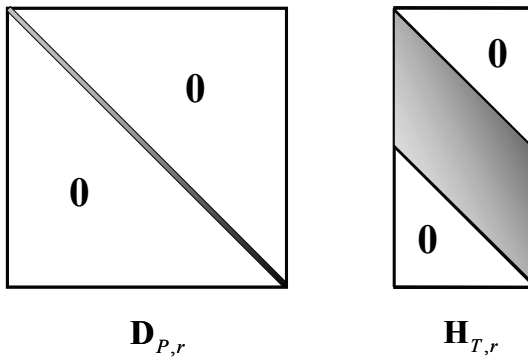


Fig. 12. Structure of (left) $\mathbf{D}_{P,r}$ and (right) $\mathbf{H}_{T,r}$ matrix. Blank parts are all 0's.

The matrix \mathbf{H}_r is a $P \times P$ lower triangular matrix with its first column vector is $[h_{1,r}, \dots, h_{L,r}, 0 \dots 0]^T$, and its first row vector is $[h_{1,r}, 0 \dots 0]$, and this matrix denotes the multipath channel over the r -th relay and destination link, L is the length of channel. Without loss of generality, we assumed that the channel lengths of different relay-destination links are all L . To avoid ISI, we should have $L \leq V$, and we assume $L = V$. The $\mathbf{D}_{p,r}$ is a diagonal matrix representing the residual carrier frequency error over the r -th relay and destination link and is defined in terms of its diagonal elements as $\mathbf{D}_{p,r} = \text{diag}(1, \alpha_r, \dots, \alpha_r^{p-1})$, with $\alpha_r = \exp(j2\pi\Delta q_r / N)$, $\text{diag}(\cdot)$ is diagonal matrix with main diagonal (\cdot) , and Δq_r is the normalized carrier frequency offset of r -th relay with the symbol duration of ZP-OFDM. Here, we notice that $\mathbf{H}_{T,r} = \mathbf{H}_r \mathbf{T}_{ZP}$ is a full column rank tall Toeplitz matrix, and its correlation matrix always guaranteed to be invertible. The structures of $\mathbf{D}_{p,r}$ and $\mathbf{H}_{T,r}$ can be shown as Fig. 12.

Since $\mathbf{H}_{T,r}$ relates to the linear convolution, we refer to this tall Toeplitz structure as linear structure, which assures symbol recovery (perfect detectability in the absence of noise) regardless of the channel zeros locations. The linear structure of ZP-OFDM provides a better BER performance and an easier blind channel estimation and blind symbol synchronization as well, while this is not the case for the CP-OFDM. In fact, the channel-irrespective symbol detectable property of ZP-OFDM is equivalent to claiming that ZP-OFDM enjoys maximum diversity gain. Intuitively, this can be understood as the ZP-OFDM retains the entire linear convolution of each transmitted symbol with the channel. Then, we will show how to use the linear property of $\mathbf{H}_{T,r}$ to achieve full spatial diversity in the cooperative system. Consequently, (27) can be rewritten as

$$\mathbf{y}_r = \mathbf{F}_{p,r} \mathbf{D}_{p,r} \mathbf{H}_{T,r} \mathbf{F}_{N,r}^H \mathbf{x}_r + \mathbf{n} \tag{29}$$

In this section, we consider a simple frequency division space frequency system for each relay \mathbf{x}_r , i.e., arranging transmitted symbols in different frequency bands according to the corresponding relay, as shown in the Fig. 13. By doing so, we can exploit the linear structure of ZP-OFDM to achieve the full cooperative diversity with linear receiver regardless of the existence of CFOs.

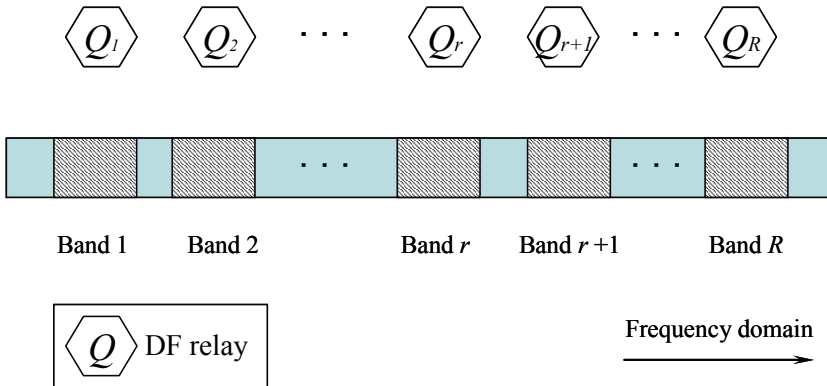


Fig. 13. Frequency division cooperative ZP-OFDM system.

We take \mathbf{x}_r as the information symbols correctly received at the r -th relay nodes involved in the DF-cooperative scheme. After full decoding, \mathbf{x}_r is assigned to the corresponding r -th frequency band as shown in the Fig. 13, and forwarded to the destination.

Considering the frequency division system, the received signal at the destination of all R relay nodes yields

$$\mathbf{y} = \mathbf{F}_p \mathbf{D} \mathbf{H} \mathbf{F}_N^H \mathbf{x} + \mathbf{n} \quad (30)$$

where $\mathbf{F}_p = \text{diag}(\mathbf{F}_{p,1}, \mathbf{F}_{p,2}, \dots, \mathbf{F}_{p,R})$, $\mathbf{D} = \text{diag}(\mathbf{D}_{p,1}, \mathbf{D}_{p,2}, \dots, \mathbf{D}_{p,R})$, $\mathbf{H} = \text{diag}(\mathbf{H}_{T,1}, \mathbf{H}_{T,2}, \dots, \mathbf{H}_{T,R})$, $\mathbf{F}_N^H = \text{diag}(\mathbf{F}_{N,1}^H, \mathbf{F}_{N,2}^H, \dots, \mathbf{F}_{N,R}^H)$ are all diagonal matrices with each relay's components on their diagonals. For instance, we consider a 2-relay cooperation system, i.e., $R = 2$, the structures of \mathbf{F}_p , \mathbf{D} , \mathbf{H} and \mathbf{F}_N^H can be illustrated as Fig. 14. In (30), $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_R^T]^T$ denotes the forwarded signal from all relays occupying different frequency bands.

If we denote $\mathbb{H} = \mathbf{F}_p \mathbf{D} \mathbf{H} \mathbf{F}_N^H$, then (30) becomes

$$\mathbf{y} = \mathbb{H} \mathbf{x} + \mathbf{n} \quad (31)$$

On the other hand, let us go back to (27), right multiplying \mathbf{T}_{ZP} changes \mathbf{H}_r from a $P \times P$ lower triangular matrix into a $P \times N$ tall Toeplitz matrix $\mathbf{H}_{T,r}$ with its first column vector as $[h_{1,r}, \dots, h_{L,r}, 0 \dots 0]^T$, and we denote $\mathbf{x}_{t,r} = \mathbf{F}_{N,r}^H \mathbf{x}_r$, which is well known as time domain signal in OFDM system, then (27) can be represented as

$$\mathbf{y}_r = \mathbf{F}_{p,r} \mathbf{D}_{p,r} \mathbf{H}_{T,r} \mathbf{x}_{t,r} + \mathbf{n} \quad (32)$$

where $\mathbf{H}_{T,r} \mathbf{x}_{t,r}$ stands for the linear convolution of the multipath channel with the time domain transmitted signal, this is a special property possessed by the ZP-OFDM system. According to the commutativity of the linear convolution, we have $\mathbf{H}_{T,r} \mathbf{x}_{t,r} = \mathbf{X}_{T,r} \mathbf{h}_r$, where $\mathbf{X}_{T,r}$ is a $P \times L$ tall Toeplitz matrix with $[\mathbf{x}_{t,r}^T, \mathbf{0}^T]^T$ as its first column, and $\mathbf{h}_r = [h_{1,r}, \dots, h_{L,r}]^T$. Consequently, (32) can be transformed into another form as

$$\mathbf{y}_r = \mathbf{F}_{p,r} \mathbf{D}_{p,r} \mathbf{X}_{T,r} \mathbf{h}_r + \mathbf{n} \quad (33)$$

We denote $\mathbf{h}_c = [\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_R^T]^T$, $\mathbf{S}_r = \mathbf{F}_{p,r} \mathbf{D}_{p,r} \mathbf{X}_{T,r}$, $\mathbb{S} = \text{diag}(\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_R)$, and consider the received signal of all R relay nodes, then we get the received signal as

$$\mathbf{y} = \mathbb{S} \mathbf{h}_c + \mathbf{n} \quad (34)$$

Equations (31) and (34) are two equivalent received data models of this frequency division cooperative ZP-OFDM system. \mathbb{H} in (31) is regarded as the overall equivalent channel, while \mathbb{S} in (34) is the equivalent signal matrix of this frequency division cooperative ZP-OFDM system. In the following section, we will exploit \mathbb{H} and \mathbf{h}_c from (31) and (34) to show the verification of the full cooperative spatial diversity.

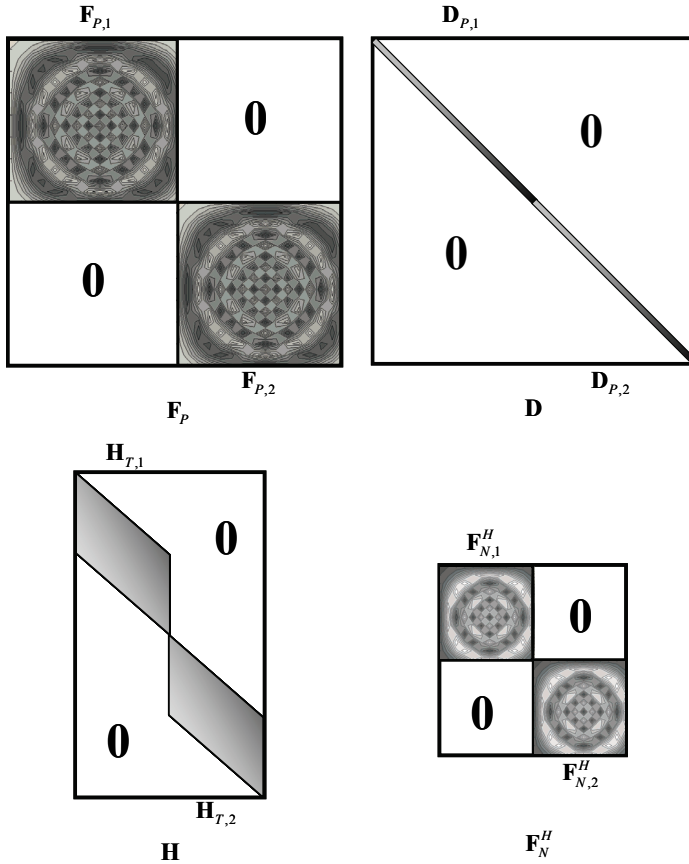


Fig. 14. Structures of the FFT matrices, CFOs matrix and channel matrix for 2-relay cooperative system, top left: FFT matrix \mathbf{F}_p , top right: CFOs matrix \mathbf{D} , bottom left: channel matrix: \mathbf{H} , bottom right: FFT matrix \mathbf{F}_N^H . Blank parts are all 0's.

4.3 Space time frequency coding design

4.3.1 Conditions of the full diversity with linear equalizer

In this section, we will show how linear receiver is the only required to achieve full cooperative diversity order RL . We first cite the following theorem from (Shang & Xia, 2007; Shang & Xia, 2008):

Theorem 2 (Shang & Xia, 2007; Shang & Xia, 2008): For PAM, PSK and square QAM constellations, if the following condition holds

$$\|\mathbb{H}\| \leq \alpha \|\mathbf{h}_c\| \quad \text{and} \quad \det(\mathbb{H}^H \mathbb{H}) \geq \beta \|\mathbf{h}_c\|^{2N}$$

where α and β are positive constants independent of \mathbf{h}_c , $\|\cdot\|$ is the Frobenius norm of a vector/matrix, and N is the number of symbols in the transmitted signal, i.e., the length of

\mathbf{x}_r in (29). Then, for any realization of \mathbb{H} , with ZF or MMSE receiver, full diversity can be achieved, i.e., the symbol error probability (SEP) P_e satisfies:

$$P_e(\hat{s}_l \rightarrow s_l) \leq \bar{c} \cdot \rho^{-RL}, \quad l=1,2,\dots,L$$

where $\bar{c} \triangleq \frac{\eta-1}{\eta}(a\hat{c})^{-2}$, η is the cardinality of the constellation, and $a = \frac{3}{2(\eta^2-1)}$,

$$\frac{\sin^2(\pi/\eta)}{2} \text{ and } \frac{3}{4(\eta-1)} \text{ for PAM, PSK, and square QAM, respectively. } \hat{c} = \beta \left(\frac{L-1}{\alpha^2} \right)^{L-1},$$

and ρ is the symbol SNR.

In what follows, we will show that, with the STFC, the frequency division cooperative ZP-OFDM system satisfies the conditions in Theorem 2. In other words, the proposed frequency division system can achieve full diversity with linear receivers. Note that here the full diversity order is RL .

4.3.2 Space time frequency coding design and verification

According to the above mentioned conditions, we design a linear structure STFC, which guarantees the full cooperative spatial diversity and without redundant power gains. By right multiplying a matrix on $\bar{\mathbf{F}}_N^H \bar{\mathbf{x}}$, where \mathbf{I}_r is an $N \times N$ identity matrix, $r \in [1, 2, \dots, R]$,

$\bar{\mathbf{F}}_N^H = \mathbf{F}_{N,r}^H$ is an N -point IFFT matrix, and $\bar{\mathbf{x}} = \mathbf{x}_r \triangleq [x_0, \dots, x_{N-1}]^T$ stands for the frequency transmitted information signal, the received signal at the destination from all R relay nodes yields

$$\mathbf{y} = \mathbf{F}_p \mathbf{D} \mathbf{H} \mathbf{G} \bar{\mathbf{F}}_N^H \bar{\mathbf{x}} + \mathbf{n} \quad (35)$$

where \mathbf{F}_p , \mathbf{D} and \mathbf{H} are the same as (30). We denote $\mathbf{H} \mathbf{G} = \hat{\mathbf{H}}_T$ and $\bar{\mathbf{x}}_t = \bar{\mathbf{F}}_N^H \bar{\mathbf{x}}$, where $\hat{\mathbf{H}}_T = [\mathbf{H}_{T,1}^T, \mathbf{H}_{T,2}^T, \dots, \mathbf{H}_{T,R}^T]^T$, $\bar{\mathbf{x}}_t$ is the time domain signal. We notice that matrix \mathbf{G} spreads $\bar{\mathbf{x}}$ according to the corresponding relays, and forms a frequency division system, since the relays perform the forwarding in the different bands, as shown clearly in the Fig.13. Therefore, the Matrix \mathbf{G} can be regarded as a coding scheme on the time domain signal, for different relays and different bands, and so called space time frequency code. Then, (35) becomes

$$\mathbf{y} = \mathbf{F}_p \mathbf{D} \hat{\mathbf{H}}_T \bar{\mathbf{F}}_N^H \bar{\mathbf{x}} + \mathbf{n} \quad (36)$$

If we denote $\mathbb{H} = \mathbf{F}_p \mathbf{D} \hat{\mathbf{H}}_T \bar{\mathbf{F}}_N^H$ as the equivalent channel matrix, we get

$$\mathbf{y} = \mathbb{H} \bar{\mathbf{x}} + \mathbf{n} \quad (37)$$

We notice that $\hat{\mathbf{H}}_T$ is a linear Toeplitz matrix. Similar to (32) and (33), we have $\hat{\mathbf{H}}_T \bar{\mathbf{x}}_t = \hat{\mathbf{X}}_T \hat{\mathbf{h}}$, where $\hat{\mathbf{X}}_T$ is a $[P \times R] \times [P \times (R-1) + L]$ tall Toeplitz matrix with $[\bar{\mathbf{x}}_t^T, \mathbf{0}^T]^T$ as its first

column, and $\hat{\mathbf{h}} = [h_{1,1}, \dots, h_{L,1}, \mathbf{0}, h_{1,2}, \dots, h_{L,2}, \mathbf{0}, \dots, h_{1,R}, \dots, h_{L,R}]^T$ with length $[P \times (R-1) + L]$, Consequently, the equivalent STFC signal matrix could be formulated as

$$\mathbf{y} = \mathbb{S} \hat{\mathbf{h}} + \mathbf{n} \quad (38)$$

where $\mathbb{S} = \mathbf{F}_p \mathbf{D} \hat{\mathbf{X}}_T$.

Its proof is shown as follows. For the first condition, we have

$$\|\mathbb{H}\| \leq \|\mathbf{F}_p\| \|\mathbf{D}\| \|\hat{\mathbf{H}}_T\| \|\bar{\mathbf{F}}_N^H\| = \|\mathbf{F}_p\| \|\mathbf{D}\| \sqrt{N} \|\hat{\mathbf{h}}\| \|\bar{\mathbf{F}}_N^H\|$$

If we take $\alpha = \sqrt{N} \|\mathbf{F}_p\| \|\mathbf{D}\| \|\bar{\mathbf{F}}_N^H\|$, we have $\|\mathbb{H}\| \leq \alpha \|\hat{\mathbf{h}}\|$, and the first condition holds.

For the second condition, to continue the proof, we cite the results from (Zhang et al., 2005) in the following theorem:

Theorem 3 (Zhang et al., 2005): For any given positive integers p and q , there exists a positive constant c , such that for any nonzero vector \mathbf{v} , $\det(\mathcal{T}^H(\mathbf{v}, p, q) \mathcal{T}(\mathbf{v}, p, q)) \geq c \|\mathbf{v}\|^{2q}$ holds.

Here, we denote $\mathcal{T}(\mathbf{v}, p, q)$ as a Toeplitz matrix of size $(p+q-1) \times q$ as follows:

$$\mathcal{T}(\mathbf{v}, p, q) \triangleq \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ v_2 & v_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_p & v_{p-1} & \cdots & 0 \\ 0 & v_p & \cdots & v_1 \\ \vdots & 0 & \cdots & v_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & v_p \end{bmatrix}$$

where $\mathbf{v} = [v_1, v_2, \dots, v_p]^T$ is any non-zeros column vector of length p .

Consequently, note that \mathbf{F}_p , \mathbf{D}_p and $\bar{\mathbf{F}}_N^H$ are all unitary matrices, we have

$$\begin{aligned} \det(\mathbb{H}^H \mathbb{H}) &= \det(\bar{\mathbf{F}}_N) \det(\bar{\mathbf{F}}_N^H) \det(\hat{\mathbf{H}}_T^H \hat{\mathbf{H}}_T) \\ &= \det(\hat{\mathbf{H}}_T^H \hat{\mathbf{H}}_T) \end{aligned}$$

where $\det(\bar{\mathbf{F}}_N) \det(\bar{\mathbf{F}}_N^H) = 1$, and $\hat{\mathbf{H}}_T$ is a tall Toeplitz matrix. According to the Theorem 3, it is easy to show

$$\det(\mathbb{H}^H \mathbb{H}) = \det(\hat{\mathbf{H}}_T^H \hat{\mathbf{H}}_T) \geq \beta \|\hat{\mathbf{h}}\|^{2N}$$

where β is a positive constant as shown in Theorem 2, and the second condition of theorem 2 holds. Therefore, the proposed STFC can achieve the full diversity for the cooperative ZP-OFDM system with CFOs and multipath channel.

According to the above full diversity scheme, we notice that the tall Toeplitz structure of the ZP-OFDM channel matrix is a unique advantage, which is not possessed by CP-OFDM. How to design a space time code or space frequency code to obtain the tall Toeplitz structure or linear convolutional structure is the key element to achieve full diversity in a cooperative wireless system.

4.4 Simulation results

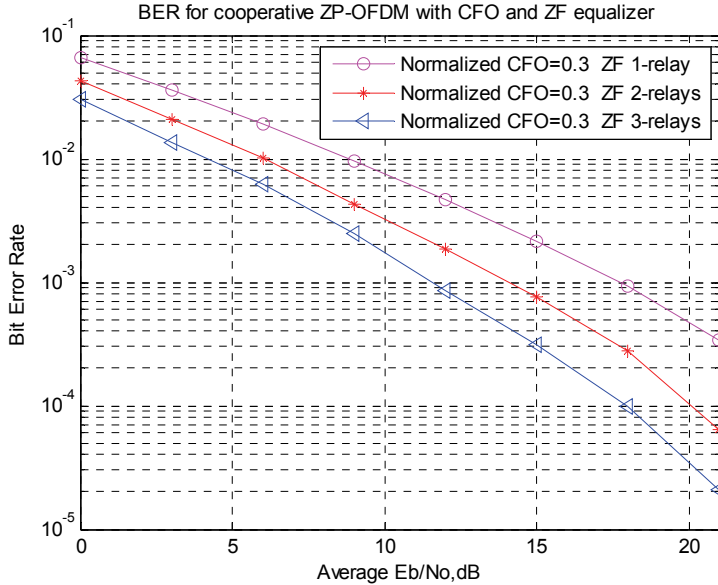


Fig. 15. BER performance of STFC cooperative ZP-OFDM with the same CFO and ZF equalizer.

Now, we present simulation results of the performance of our STFC design. For different numbers of relays with ZF equalizer, Fig. 15 shows the BER vs. E_b/N_0 performance under the same CFO, and $\Delta q_r = 0.3$. The diversity order can be shown as the asymptotic slope of BER vs. E_b/N_0 curve. It describes how fast the error probability decays with SNR. We can see from Fig. 15 that the full cooperative diversity is achieved with the ZF equalizer, because the asymptotic slope of the curve increases with the increasing of the number of relays.

Fig. 16 shows the BER vs. E_b/N_0 performance with different CFOs and with MMSE equalizer. We assume that the absolute value of CFOs is less than half of the subcarrier spacing, i.e., $\Delta q_r \in (-0.5, 0.5)$. It can be seen that, the E_b/N_0 gaps between $\Delta q_r = 0.01$ and $\Delta q_r = 0.5$ are very small. Although the increasing of CFOs will not change the diversity orders, it will deteriorate the BER performance because larger CFOs result in larger approximation error. In other words, larger CFOs will increase the power of equivalent noise and consequently degrade the communication performance. We can also see that the full cooperative diversity is achieved with the MMSE equalizers. Compared to ZF equalizer, the BER performance of MMSE equalizer is much better. In the Fig. 16, even with the severer CFOs, diversity increases with increasing the number of relays. From the figure, we also

notice that, without the proposed STFC, i.e., direct combining of the 2-relay signals at the destination, the 2-relays system only yields 3 dB power gain. The 2-relay BER vs. E_b/N_0 curve without STFC is parallel to the 1-relay case, indicating no diversity is achieved.

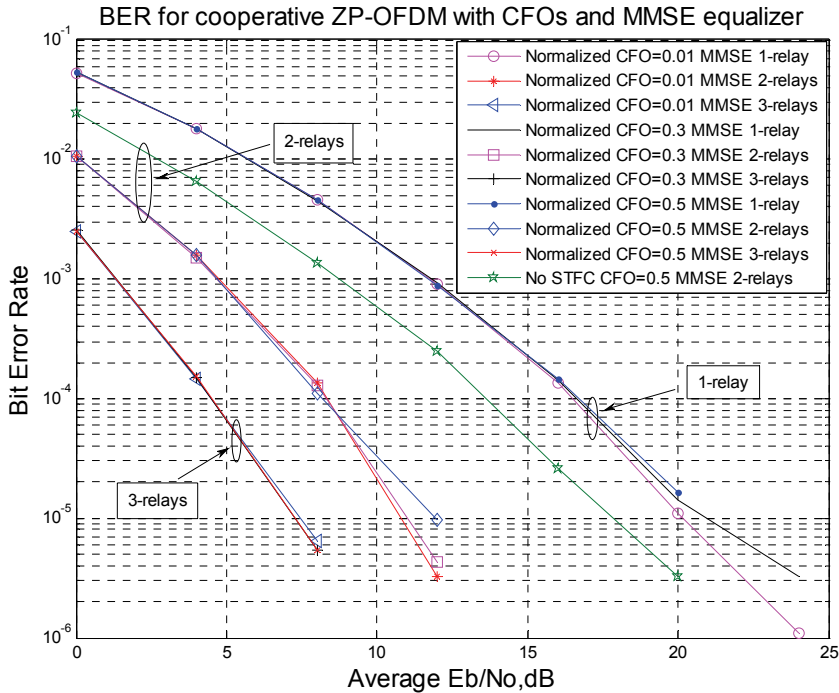


Fig. 16. BER performance of linear freq. div. STFC cooperative ZP-OFDM with the different CFOs and MMSE equalizer.

5. Conclusions

In this chapter, we investigated some cooperative OFDM communication issues, such as relay selection and full diversity design. First, we proposed a hybrid OFDM cooperative strategy for multi-node wireless networks employing both DF and AF relaying. Fully decoding is guaranteed by simply comparing SNRs at relay nodes to the SNR threshold, which is more efficient than utilizing conventional cyclic redundant checking code. The lower bound and the upper bound of the SNR threshold were provided as well. After correct decoding, the DF protocol outperforms AF protocol in terms of BER performance, which can be seen from the Monte Carlo simulation as well as analytical results. These results justify that the DF protocol dominates hybrid cooperation strategy. For the suggested hybrid DF-AF cooperation protocol, we also represented a dynamic optimal combination strategy for the optimal AF selection. The closed-form BER expression of the hybrid OFDM cooperation in Rayleigh fading channel was derived. The agreement between the analytical curves and numerical simulated results shows that the derived closed-form BER expression is suitable for the DF-dominant hybrid cooperation protocols. The compact

and closed-form BER expression can easily provide an insight into the results as well as a heuristic help for the design of future cooperative wireless systems. Subsequently, we investigated the cooperative ZP-OFDM communication with multiple CFOs and multipath channel, i.e., with ICI and ISI. In the ZP-OFDM system, the linear structure, or tall Toeplitz structure of channel matrix is a unique feature. Therefore, we provided a simple frequency division cooperative ZP-OFDM system to illustrate the full diversity design, and showed that some power expenditure was needed for the proposed frequency division cooperative ZP-OFDM system to achieve full cooperative diversity. Then, we proposed another power efficient STFC, taking advantage of the linear structure of ZP-OFDM, to achieve the full cooperative diversity. Furthermore, we also showed that, linear equalizations, such as ZF and MMSE equalizers, can be used to collect the full cooperative spatial diversity gain. Comparing to exhausted ML decoding methods, the computational complexity is greatly reduced.

6. References

- Lima, P.; Bonarini, A. & Mataric, M. (2004). *Name of Book in Italics*, Publisher, ISBN, Place of Publication
- Batra, A. Multi-band OFDM physical layer proposal for IEEE 802.15Task Group 3a IEEE P802.15-04/0493r1, Sep. 2004.
- Batra, A.; Balakrishnan, J.; Aiello, G. R. & Dabak, A. (2004). Design of a multiband OFDM system for realistic UWB channel environments, *IEEE Transaction on Microwave and Techniques*, Vol. 52, No. 9, pp. 2123–2138, ISSN: 0018-9480
- Barbarossa, S. (2005). *Multiantenna Wireless Communication Systems*, MA: Artech House, ISBN-13: 978-1580536349, Norwood
- Cover, T. M. & El Gamal, A. A. (1979). Capacity theorems for the relay channel, *IEEE Trans. Inform. Theory*, Vol. 25, No. 5, pp. 572-584, ISSN: 0018-9448
- Farhadi, G. & Beaulieu, N. C. (2008). On the Ergodic Capacity of Wireless Relaying System over Rayleigh Fading Channels, *IEEE Trans. Wireless Communications*, Vol. 7, No. 11, pp. 4462-4467, ISSN: 1536-1276
- Foschini, G. J. & Gans, M. (1998). On the limits of wireless communication in a fading environment when using multiple antennas, *Wireless Personal Commun.*, Vol. 6, pp. 311-335, ISSN: 0929-6212
- Hasna, M. O. & Alouini, M. -S. (2002). Performance analysis of two-hop relayed transmissions over Rayleigh-fading channels, *Proceedings of Vehicular Technology Conf.* 2002, pp. 1992-1996, Birmingham, AL
- Hwang, J.; Consulta, R. R. & Yoon, H. (2007) 4G mobile networks - technology beyond 2.5G and 3G. In *PTC*, 2007.
- Kramer, G.; Gastpar, M. & Gupta, P. (2005). Cooperative strategies and capacity theorems for relay networks, *IEEE Trans. Inf. Theory*, Vol. 51, pp. 3037–3063, ISSN: 0018-9448
- Laneman, J. N.; Tse, D. N. C. & Wornell, G. W. (2004). Cooperative diversity in wireless networks: efficient protocols and outage behavior. *IEEE Trans. Inform. Theory*, Vol. 50, No. 12, pp. 3062-3080, ISSN: 0018-9448
- Li, Y.; Zhang, W. & Xia, X. (2009). Distributive high rate space-frequency codes achieving full cooperative and multipath diversities for asynchronous cooperative communications, *IEEE Trans. Vehicular Technology*, Vol. 58, No. 1, pp. 207-217, ISSN: 0018-9545

- Lin, S. & Costello, D. J. Jr. (1983). *Error Control Coding: Fundamentals and Applications.*, NJ: Prentice-Hall, ISBN: 013283796X, Englewood Cliffs
- Liu, Z.; Xin, Y. & Giannakis, G. B. (2003). Linear constellation precoded OFDM with maximum multipath diversity and coding gains, *IEEE Trans. Commun.*, Vol. 51, No. 3, pp. 416–427, ISSN: 0090-6778
- Liu, K. J. R.; Sadek, A. K.; Su W. & Kwasinski, A. (2009). *Cooperative communications and networks*, Cambridge University Press, ISBN-13 978-0-521-89513-2, Cambridge
- Louie, R.; Li, Y.; Suraweera, H. A. & Vucetic, B. (2009). Performance analysis of beamforming in two hop amplify and forward relay network with antenna correlation, *IEEE Trans. Wireless Communications*, Vol. 8, No. 6, pp. 3132-3141, ISSN: 1536-1276
- Lu, H.; Nikoogar, H. & Lian, X. (2010). Performance evaluation of hybrid DF-AF OFDM cooperation in Rayleigh Channel, to appear in European Wireless Technology Conference, 2010
- Lu, H. & Nikoogar, H. (2009). A thresholding strategy for DF-AF hybrid cooperative wireless networks and its performance, Proceedings of IEEE SCVT '09, UCL, Louvain. Nov. 2009
- Lu, H.; Nikoogar, H. & Chen, H. (2009). On the potential of ZP-OFDM for cognitive radio, Proceedings of WPMC'09, pp. 7-10, Sendai, Japan. Sep. 2009
- Ma, X. & Zhang, W. (2008). Fundamental limits of linear equalizers: diversity, capacity, and complexity, *IEEE Trans. Inform. Theory*, Vol. 54, No. 8, pp. 3442-3456, ISSN: 0018-9448
- Muquet, B.; Wang, Z.; Giannakis, G. B.; Courville, M. & Duhamel, P. (2002). Cyclic prefixing or zero padding for wireless multicarrier transmissions, *IEEE Transaction on Communications*, Vol. 50, No. 12, pp. 2136–2148, ISSN: 0090-6778
- Nosratinia, A.; Hunter, T. E. & Hedayat, A. (2004). Cooperative communication in wireless networks, *IEEE Commun. Mag.*, Vol. 42, pp. 74–80, ISSN: 0163-6804
- Patel, C.; Stüber, G. & Pratt, T. (2006). Statistical properties of amplify and forward relay fading channels, *IEEE Trans. Veh. Technol.*, Vol. 55, No. 1, pp. 1-9, ISSN: 0018-9545
- Proakis, J. G. (2001). *Digital Communications*, 4th ed., McGraw Hill, ISBN-13: 978-0072321111, New York
- Sadek, A. K.; Su, W. & Liu, K. J. R. (2007). Multi-node cooperative communications in wireless networks, *IEEE Trans. Signal Processing*, Vol. 55, No. 1, pp. 341-355, ISSN: 1053-587X
- Sendonaris, A.; Erkip, E. & Aazhang, B. (2003). User cooperation diversity – Part I: System description, *IEEE Trans. Commun.*, Vol. 51, No. 11, pp. 1927–1938, ISSN: 0090-6778
- Sendonaris, A.; Erkip, E. & Aazhang, B. (2003). User cooperation diversity – Part II: Implementation aspects and performance analysis, *IEEE Trans. Commun.*, Vol. 51, No. 11, pp. 1939–1948, ISSN: 0090-6778
- Shang, Y. & Xia, X.-G. (2007). A criterion and design for space-time block codes achieving full diversity with linear receivers, Proceedings of IEEE ISIT'07, Nice, France, pp. 2906–2910, June 2007
- Shang, Y. & Xia, X.-G. (2008). On space-time block codes achieving full diversity with linear receivers, *IEEE Trans. Inform. Theory*, Vol. 54, pp. 4528–4547, ISSN: 0018-9448
- Standard ECMA-368 High Rate Ultra Wideband PHY and MAC Standard, 3rd edition, Dec. 2008

- Tse, D. N. C. & Viswanath, P. (2005). *Fundamentals of Wireless Communications.*, U.K.: Cambridge Univ. Press, ISBN-13: 9780521845274, Cambridge
- Wang, Z. & Giannakis, G. B. (2000). Wireless multicarrier communications: Where Fourier meets Shannon, *IEEE Signal Processing Mag.*, Vol. 17, No. 3, pp. 1-17, ISSN: 1053-5888
- Zhang, J.-K.; Liu, J. & Wong, K. M. (2005). Linear Toeplitz space time block codes, Proceedings of IEEE ISIT'05, Adelaide, Australia, Sept. 2005

High Throughput Transmissions in OFDM based Random Access Wireless Networks

Nuno Souto^{1,2}, Rui Dinis^{2,3}, João Carlos Silva^{1,2},
Paulo Carvalho³ and Alexandre Lourenço^{1,2}

¹ISCTE-IUL

²Instituto de Telecomunicações,

³UNINOVA/FCT-UNL,
Portugal,

1. Introduction

In Random Access Wireless Networks it is common to occur packet collisions due to different users trying to access simultaneously to a given physical channel. The conventional approach is to discard all blocks involved in the collision and retransmit them again. To reduce the chances of multiple collisions each user transmits in the next available slot with a given probability. With this strategy, if two packets collide we need at least three time slots to complete the transmission (more if there are multiple collisions), which results in a throughput loss.

To overcome this problem, a TA (Tree Algorithm) combined with a SIC (Successive Interference Cancellation) scheme was proposed in (Yu & Giannakis, 2005). Within that scheme, the signal associated to a collision is not discarded. Instead, if the packets of two users collide then, once we receive with success the packet of one of those users, we can subtract the corresponding signal from the signal with collision and recover the packet from the other user. With this strategy, a collision involving two packets requires only one additional time slot to complete the transmission, unless there are multiple collisions. However, the method has a setback since possible decision errors might lead to a deadlock. (Wang et al., 2005) Another problem with these techniques is that we do not take full advantage of the information in the collision. The ideal situation would be to use the signals associated to multiple collisions to separate the packets involved (in fact, solving collisions can be regarded as a multiuser detection problem). In (Tsatsanis et al., 2000) a multipacket detection technique was proposed where all users involved in a collision of N_p packets retransmit their packets N_p-1 times, each one with a different phase rotation to allow packet separation. However, this technique is only suitable for flat-fading channels (there are phase rotations that might lead to an ill-conditioned packet separation). Moreover, it is difficult to cope with channel variations during the time interval required to transmit the N_p variants of each packets (the same was also true for the SIC-TA technique of (Yu & Giannakis, 2005). A variant of these techniques suitable for time-dispersive channels was proposed in (Zhang & Tsatsanis, 2002) although the receiver complexity can become very high for severely time-dispersive channels.

A promising method for resolving multiple collisions was proposed in (Dinis, et al., 2007) for SC modulations (Single Carrier) with FDE (Frequency-Domain Equalization). Since that technique is able to cope with multiple collisions, the achievable throughputs can be very high (Dinis, et al., 2007). In this chapter we extend that approach to wireless systems employing OFDM modulations (Orthogonal Frequency Division Multiplexing) (Cimini, 1985), since they are currently being employed or considered for several digital broadcast systems and wireless networks (Nee & Prasad, 2000) (3GPP TR25.814, 2006). To detect all the simultaneously transmitted packets we propose an iterative multipacket receiver capable of extracting the packets involved in successive collisions. The receiver combines multipacket separation with interference cancellation (IC). To be effective our receiver requires uncorrelated channels for different retransmissions. Therefore, to cope with quasi-stationary channels, different interleaved versions of the data blocks are sent in different retransmissions.

In this chapter it is also given some insight into the problem of estimating the number of users involved in a collision by analyzing the probability distribution of the decision variable and selecting a convenient detection threshold. The problem of estimating the channel characteristics (namely the channel frequency response) of each user is also addressed. Regarding this issue and due to its iterative nature the proposed receiver can perform enhanced channel estimation.

The chapter is organized as follows. First the system model is defined in Section 2 while Section 3 and 4 describe the proposed transmitter and multipacket receiver in detail. The MAC scheme is analyzed in Section 5 while Section 6 presents some performance results. Finally the conclusions are given on Section 7.

2. System description

In this chapter we consider a random access wireless network employing an OFDM scheme with N subcarriers where each user can transmit a packet in a given time slot. If N_p users decide to transmit a packet in the same time slot then a collision involving N_p packets will result. In this case, all packets involved in the collision will be retransmitted N_p-1 times. In practice, the receiver (typically the BS - Base Station) just needs to inform all users of how many times they have to retransmit their packets (and in which time-slots, to avoid collisions with new users). The request for retransmissions can be implemented very simply with a feedback bit that is transmitted to all users. If it is a '1' any user can try to transmit in the next time slot. When it becomes '0' the users that tried to transmit in the last time slot must retransmit their packets in the following time slots until the bit becomes a '1'. All the other users cannot transmit any packet while the bit is '0'.

The receiver detects the packets involved in the collision as soon as it receives N_p different signals associated to the collision of the N_p packets. The figure (Fig. 1) illustrates the sequence of steps using an example with 2 users.

At the receiver, the basic idea is to use all these received transmission attempts to separate the N_p colliding packets. In fact, our system can be regarded as a MIMO system (Multiple-Input, Multiple Output) where each input corresponds to a given packet and each output corresponds to each version of the collision. To accomplish a reliable detection at the receiver it is important that the correlation between multiple received retransmissions (i.e., multiple versions of each packet involved in the collision) is as low as possible. For static or slow-varying channels this correlation might be very high, unless different frequency bands

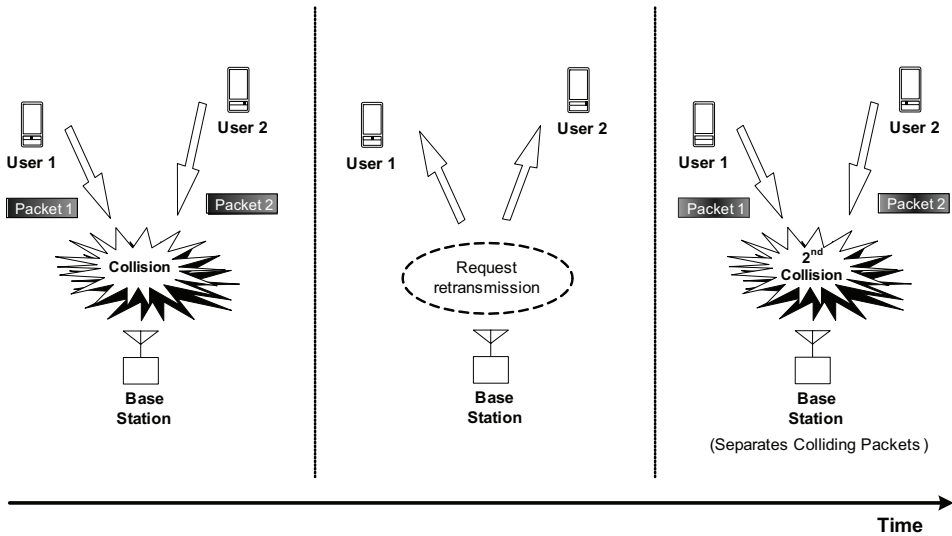


Fig. 1. Sequence of steps required for the multipacket detection method for the case of 2 colliding packets.

are adopted for each retransmission. To overcome this problem, we can take advantage of the nature of OFDM transmission over severely time-dispersive channels where the channel frequency response can change significantly after just a few subcarriers. This means that the channel frequency response for subcarriers that are not close (i.e., subcarriers in different parts of the OFDM band) can be almost uncorrelated. Therefore, by simply applying a different interleaving to the modulated symbols in each retransmission it is possible to reduce the correlation between them¹. In this chapter we will call them symbol interleavers to distinguish from the other interleaving blocks²).

3. Transmitter design

In Fig. 2 it is shown the block diagram representing the processing chain of a transmitter designed to be used with the proposed packet separation scheme.

According to the diagram the information bits are first encoded and rate matching is applied to fit the sequence into the radio frame, which is accomplished by introducing or removing bits. The resulting encoded sequence is interleaved and mapped into complex symbols according to the chosen modulation. A selector then chooses to apply a symbol interleaver or not depending on whether it is a retransmission or the first transmission attempt. A total

¹ Clearly, using different symbol-level interleavers before mapping the coded symbols in the OFDM subcarriers is formally equivalent to interleave the channel frequency response for different subcarriers. For a given subcarrier, this reduces the correlation between the channel frequency response for different retransmissions.

² It should be pointed out that in this chapter we assume that the interleaver to reduce the correlation between different retransmissions operates at the symbol level and the interleavers associated to the channel encoding are at the bit level. However, all interleavers could be performed at the bit level.

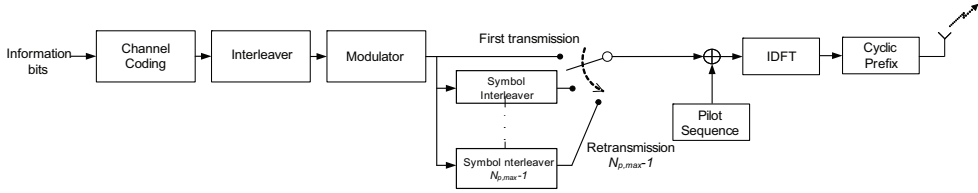


Fig. 2. Emitter Structure

of $N_{p,max}-1$ different symbol interleavers are available, where $N_{p,max}$ is the maximum number of users that can try to transmit simultaneously, so that a different one is applied in each retransmission. Known pilot symbols are inserted into the modulated symbols sequence before the conversion to the time domain using an IDFT (Inverse Discrete Fourier Transform). As will be explained further ahead, the pilot symbols are used for accomplishing user activity detection and channel estimation at the base station.

4. Receiver design

4.1 Receiver structure

To detect the multiple packets involved in a collision we propose the use of an iterative receiver whose structure is shown in Fig. 3.

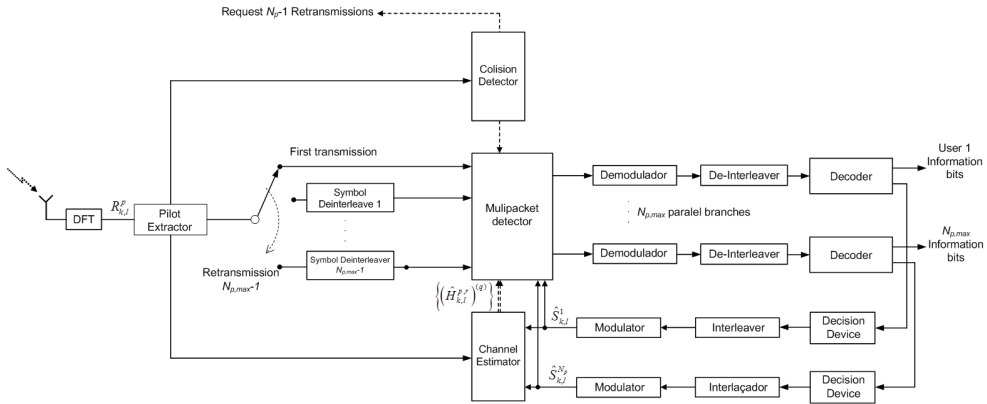


Fig. 3. Iterative receiver structure.

For simplicity we will assume that different packets arrive simultaneously. In practice, this means that some coarse time-advance mechanism is required, although some residual time synchronization error can be absorbed by the cyclic prefix. As with other OFDM-based schemes, accurate frequency synchronization is also required. First, the received signals corresponding to different retransmissions, which are considered to be sampled and with the cyclic prefix removed, are converted to the frequency domain with an appropriate size- N DFT operation. Pilot symbols are extracted for user activity detection in the "Collision Detection" block as well as for channel estimation purposes while the data symbols are de-interleaved according to the retransmission to which they belong.

Assuming that the cyclic prefix is longer than the overall channel impulse response (the typical situation in OFDM-based systems) the resulting sequence for the r th transmission attempt can be written as:

$$\mathbf{R}_{k,l}^r = \sum_{p=1}^{N_p} S_{k,l}^p H_{k,l}^{p,r} + N_{k,l}^r \quad (1)$$

with $H_{k,l}^{p,r}$ denoting the overall channel frequency response in the k^{th} frequency of the l^{th} OFDM block for user p during transmission attempt r . $N_{k,l}^p$ denotes the corresponding channel noise and $S_{k,l}^p$ is the data symbol selected from a given constellation, transmitted on the k^{th} ($k=1, \dots, N$) subcarrier of the l^{th} OFDM block by user p ($p=1, \dots, N_p$). Since we are applying interleaving to the retransmissions, to simplify the mathematical representation we will just assume that it is the sequence of channel coefficients $H_{k,l}^{p,r}$ that are interleaved instead of the symbols (therefore we do not use the index r in $S_{k,l}^p$).

After the symbol de-interleavers the sequences of samples associated to all retransmissions are used for detecting all the packets inside the Multipacket Detector with the help of a channel estimator block. After the Multipacket Detector, the demultiplexed symbols sequences pass through the demodulator, de-interleaver and channel decoder. This channel decoder has two outputs: one is the estimated information sequence and the other is the sequence of log-likelihood ratio (LLR) estimates of the code symbols. These LLRs are passed through the Decision Device which outputs soft-decision estimates of the code symbols. These estimates enter the Transmitted Signal Rebuilder which performs the same operations of the transmitters (interleaving, modulation). The reconstructed symbol sequences are then used for a refinement of the channel estimates and also for possible improvement of the multipacket detection task for the subsequent iteration. This can be accomplished using an IC in the Multipacket Detector block.

4.2 Multipacket Detector

The objective of the Multipacket Detector is to separate multiple colliding packets. It can accomplish this with several different methods. In the first receiver iterations it can apply either the MMSE criterion (Minimum Mean Squared Error), the ZF criterion (Zero Forcing) or a Maximum Likelihood Soft Output criterion (MLSO) (Souto et al., 2008). Using matrix notation the MMSE estimates of the transmitted symbols in subcarrier k and OFDM block l is given by

$$\hat{\mathbf{S}}_{k,l} = \hat{\mathbf{H}}_{k,l}^H \cdot \left(\hat{\mathbf{H}}_{k,l} \hat{\mathbf{H}}_{k,l}^H + \sigma^2 \mathbf{I} \right)^{-1} \mathbf{R}_{k,l} \quad (2)$$

where $\hat{\mathbf{S}}_{k,l}$ is the $N_p \times 1$ estimated transmitted signal vector with one user in each position, $\hat{\mathbf{H}}_{k,l}$ is the $N_p \times N_p$ channel matrix estimate with each column representing a different user and each line representing a different transmission attempt, $\mathbf{R}_{k,l}$ is the $N_p \times 1$ received signal vector with one received transmission attempt in each position and σ^2 is the noise variance. The ZF estimate can be simply obtained by setting σ to 0 in (2). In the MLSO criterion we use the following estimate for each symbol

$$\hat{S}_{k,l}^p = E \left[S_{k,l}^p \mid \mathbf{R}_{k,l} \right] = \sum_{s_i \in \Lambda} s_i \cdot \frac{P(S_{k,l}^p = s_i)}{p(\mathbf{R}_{k,l})} p(\mathbf{R}_{k,l} \mid S_{k,l}^p = s_i) \quad (3)$$

where s_i corresponds to a constellation symbol from the modulation alphabet Λ , $E[\cdot]$ is the expected value, $P(\cdot)$ represents a probability and $p(\cdot)$ a probability density function (PDF). Considering equiprobable symbols $P(S_{k,l}^p = s_i) = 1/M$, where M is the constellation size. The PDF values required in (3) can be computed as:

$$\begin{aligned}
 p(\mathbf{R}_{k,l} | S_{k,l}^p = s_i) &= \frac{1}{M^{N_p-1}} \sum_{\mathbf{S}_{k,l}^{\text{interf}} \in \Lambda^{N_p-1}} p(\mathbf{R}_{k,l} | S_{k,l}^p = s_i, \mathbf{S}_{k,l}^{\text{interf}}) \\
 &= \frac{1}{M^{N_p-1}} \sum_{\mathbf{S}_{k,l}^{\text{interf}} \in \Lambda^{N_p-1}} \frac{1}{(2\pi\sigma^2)^{N_p}} \exp \left[- \sum_{r=1}^{N_p} \frac{\left| R_{k,l}^r - \sum_{m=1}^{N_p} S_{k,l}^m \hat{H}_{k,l}^{m,r} \right|^2}{2\sigma^2} \right]
 \end{aligned} \quad (4)$$

Where $\mathbf{S}_{k,l}^{\text{interf}}$ is a $(N_p-1) \times 1$ vector representing a possible combination of colliding symbols except the one belonging to packet p . An interference canceller (IC) can also be used inside the Multipacket Detector, but usually is only recommendable after the first receiver iteration (Souto et al., 2008). In iteration q , for each packet p in each transmission attempt r , the IC subtracts the interference caused by all the other packets in that attempt. This can be represented as:

$$\left(R_{k,l}^{r,p} \right)^{(q)} = R_{k,l}^r - \sum_{\substack{m=1 \\ m \neq p}}^{N_p} \left(\hat{S}_{k,l}^m \right)^{(q-1)} \hat{H}_{k,l}^{m,r} \quad (5)$$

Where $\left(\hat{S}_{k,l}^m \right)^{(q-1)}$ is the transmitted symbol estimate obtained in the previous iteration for packet m , subcarrier k and OFDM block l .

4.3 Channel estimation

To achieve coherent detection at the receiver known pilot symbols are periodically inserted into the data stream. The proposed frame structure is shown in Fig. 4. For an OFDM system with N carriers, pilot symbols are multiplexed with data symbols using a spacing of ΔN_T OFDM blocks in the time domain and ΔN_F subcarriers in the frequency domain. To avoid interference between pilots of different users, FDM (Frequency Division Multiplexing) is employed for the pilots, which means that pilot symbols cannot be transmitted over the same subcarrier by different users. No user can transmit data symbols on subcarriers reserved for pilots, therefore, the minimum allowed spacing in the frequency domain is $(\Delta N_F)_{\min} = N_{p,\max}$, where $N_{p,\max}$ is the maximum number of users that can try to transmit simultaneously.

To obtain the frequency channel response estimates for each transmitting/receiving antenna pair the receiver applies the following steps in each iteration:

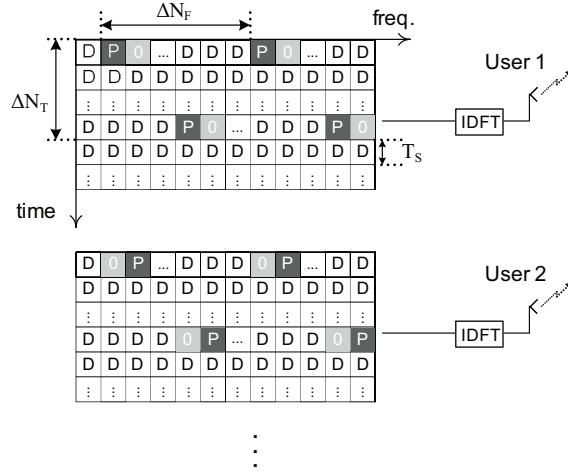


Fig. 4. Proposed frame structure for MIMO-OFDM transmission with implicit pilots (P – pilot symbol, D – data symbol, 0 – empty subcarrier).

1. The channel estimate between transmit antenna m and receive antenna n for each pilot symbol position, is simply computed as:

$$\tilde{H}_{k,l}^{p,r} = \frac{\left(S_{k,l}^{p,Pilot}\right)^*}{\left|S_{k,l}^{p,Pilot}\right|^2} R_{k,l}^r \quad (6)$$

where $S_{k,l}^{p,Pilot}$ corresponds to a pilot symbol transmitted in the k^{th} subcarrier of the l^{th} OFDM block by user p . Obviously, not all indexes k and l will correspond to a pilot symbol since $\Delta N_T > 1$ or $\Delta N_F > 1$.

2. Channel estimates for the same subcarrier k , user p and transmission attempt r but in time domain positions (index l) that do not carry a pilot symbol can be obtained through interpolation using a finite impulse response (FIR) filter with length W as follows:

$$\tilde{H}_{k,l+t}^{p,r} = \sum_{j=-\lfloor(W-1)/2\rfloor}^{\lfloor W/2\rfloor} h_t^j \tilde{H}_{k,l+j\Delta N_T}^{p,r} \quad (7)$$

where t is the OFDM block index relative to the last one carrying a pilot (which is block with index l) and h_t^j are the interpolation coefficients of the estimation filter which depend on the channel estimation algorithm employed. There are several proposed algorithms in the literature like the optimal Wiener filter interpolator (Cavers, 1991) or the low pass sinc interpolator (Kim et al., 1997).

3. After the first iteration the data estimates can also be used as pilots for channel estimation refinement (Valenti, 2001). The respective channel estimates are computed as

$$\left(\tilde{H}_{k,l}^{p,r}\right)^{(q)} = \frac{R_{k,l}^r \left(\hat{S}_{k,l}^p\right)^{(q-1)*}}{\left|\left(\hat{S}_{k,l}^p\right)^{(q-1)}\right|^2} \quad (8)$$

4. The channel estimates are enhanced by ensuring that the corresponding impulse response has a duration N_G (number of samples at the cyclic prefix). This is accomplished by computing the time domain impulse response through $\left\{\left(\tilde{h}_{i,l}^{p,r}\right)^{(q)} ; i=0,1,\dots,N-1\right\} = \text{DFT}\left\{\left(\tilde{H}_{k,l}^{p,r}\right)^{(q)} ; k=0,1, \dots,N-1\right\}$, followed by the truncation of this sequence according to $\left\{\left(\hat{h}_{i,l}^{p,r}\right)^{(q)} = w_i \left(\tilde{h}_{i,l}^{p,r}\right)^{(q)} ; i=0,1,\dots,N-1\right\}$ with $w_i = 1$ if the i^{th} time domain sample is inside the cyclic prefix duration and $w_i = 0$ otherwise. The final frequency response estimates are then obtained as $\left\{\left(\hat{H}_{k,l}^{p,r}\right)^{(q)} ; k=0,1,\dots,N-1\right\} = \text{IDFT}\left\{\left(\hat{h}_{i,l}^{p,r}\right)^{(q)} ; i=0,1,\dots,N-1\right\}$.

4.4 Detection of users involved in a collision

One of the difficulties of employing multipacket detector schemes, namely the ones proposed in this chapter, lies in finding out which users have packets involved in the collision. Missing a user will result in an insufficient number of retransmissions to reliably extract the others while assuming a non-transmitting user as being active will also degrade the packet separation and waste resources by requesting an excessive number of retransmissions. In the following we propose a simple detection method that can be combined with the multipacket detection approach described previously. This method considers the use of OFDM blocks with pilots multiplexed with conventional data blocks, as described in the previous subsection. We assume that the maximum number of users that can attempt to transmit their packets in a given physical channel is $N_{p,\max}$. Since each user p has a specific subset of subcarriers reserved for its pilot symbols the receiver can use those subcarriers to estimate whether the user is transmitting a packet or not. To accomplish that objective it starts by computing the decision variable:

$$Y_p = \sum_{k',l'}^{N_{pilots}} \left|R_{k',l'}^1\right|^2, \quad p = 1,\dots,N_{p,\max} \quad (9)$$

for all users, with (k',l') representing all positions (subcarriers and OFDM blocks) containing a pilot symbol of user p and N_{pilots} being the total number of pilots used inside the sum. The decision variable, Y_p , can then be compared with a threshold y^{th} to decide if a user is active or not.

The threshold should be chosen so as to maximize the overall system throughput. Assuming a worst-case scenario where any incorrect detection of the number of users results in the loss of all packets then, from (Tsatsanis et al., 2000), the gross simplified system throughput (not taking into account bit errors in decoded packets) is given by:

$$R = \frac{N_{p,\max}(1-P_e)}{N_{p,\max}(1-P_e) + P_e N_{p,\max}} (1-P_M) \left[(1-P_e)(1-P_M) + P_e(1-P_F) \right]^{N_{p,\max}-1} \quad (10)$$

where P_e is the probability of a user's buffer being empty at the beginning of a transmission slot, P_M is the probability of a missed detection and P_F is the false alarm probability. The threshold, y^{th} , that maximizes (10) can be found through:

$$\frac{\partial R}{\partial y} = 0 \quad (11)$$

resulting

$$\frac{\partial P_M}{\partial y} \left[(1 - P_e)(1 - P_M)N_{p,\max} + P_e(1 - P_F) \right] = (N_{p,\max} - 1)(1 - P_M)P_e \frac{\partial(1 - P_F)}{\partial y} \quad (12)$$

Assuming low false alarm and missed detection probabilities, i.e.,

$$\begin{cases} 1 - P_M \approx 1 \\ 1 - P_F \approx 1 \end{cases} \quad (13)$$

and noting that:

$$\begin{cases} p_1(y) = \frac{\partial(1 - P_F)}{\partial y} \\ p_2(y) = \frac{\partial P_M}{\partial y} \end{cases} \quad (14)$$

where $p_1(y)$ is the probability density function (PDF) of $\sum_{k',l'}^{N_{pilots}} |N_{k',l'}^1|^2$ and $p_2(y)$ is the PDF of $\sum_{k',l'}^{N_{pilots}} \left| S_{k',l'}^{p,pilot} H_{k',l'}^{p,1} + N_{k',l'}^1 \right|^2$, (12) can be rewritten as

$$p_1(y) = p_2(y) \frac{N_{p,\max} - P_e(N_{p,\max} - 1)}{(N_{p,\max} - 1)P_e} \quad (15)$$

Therefore we can compute the threshold from the weighted intersection of the two PDFs, $p_1(y)$ and $p_2(y)$. Regarding the first PDF, since $N_{k,l}^1$ are zero mean independent complex Gaussian variables with variance $E\left[|N_{k,l}^1|^2\right] = N_0$ ($N_0/2$ is the noise power spectral density), $|N_{k,l}^1|^2$ will have an exponential distribution with average $\mu_1 = E\left[|N_{k,l}^1|^2\right]$.

Therefore the decision variable corresponds to a sum of independent exponential random variables and, as a result, follows an Erlang distribution expressed as

$$p_1(y) = \frac{y^{N_{pilots}-1} \exp\left(-\frac{y}{\mu_1}\right)}{\mu_1^{N_{pilots}} (N_{pilots} - 1)!} \quad (16)$$

Regarding the second PDF, $R_{k,l}^1 = S_{k,l}^{pilot} H_{k,l}^{p,1} + N_{k,l}^1$ and $|R_{k,l}^1|^2$ are also zero mean complex

Gaussian and exponential variables with average given by $\mu_2 = |S_{k,l}^{pilot}|^2 E[|H_{k,l}^{p,1}|^2] + N_0$,

respectively. However they are not necessarily uncorrelated for different k and l . Since the receiver does not have a priori knowledge about the PDP (Power Delay Profile) of each user while it is still detecting them it does not know the correlation between different channel frequency response coefficients. For that reason, we opted to employ a threshold located in the middle of those obtained assuming two extreme cases: uncorrelated channel frequency response coefficients and constant channel frequency response coefficients.

4.5 Uncorrelated channel frequency response

If the different channel frequency response coefficients, $H_{k,l}^{p,1}$, can be assumed uncorrelated for different k and l (for example a severe time-dispersive channel) then the decision variable Y_p will correspond to a sum of uncorrelated exponential variables resulting again in an Erlang random variable described by the following PDF

$$p_2(y) = \frac{y^{N_{pilots}-1} \exp\left(-\frac{y}{\mu_2}\right)}{\mu_2^{N_{pilots}} (N_{pilots}-1)!} \quad (17)$$

Therefore, the intersection of PDFs (16) and (17) results in the threshold given by

$$y^{th} = \frac{N_{pilots} \ln\left(\frac{\mu_2}{\mu_1}\right)}{\frac{1}{\mu_1} - \frac{1}{\mu_2}} \quad (18)$$

4.6 Constant channel frequency response

If the channel is basically non time dispersive then the channel frequency response coefficients, $H_{k,l}^{p,1}$, will be almost constant for different k and l and, thus, the decision variable Y_p will correspond to a sum of correlated exponential variables. To obtain the PDF for this case it is necessary to remind the fact that the exponential distribution is a special case of the gamma distribution. Consequently, we can employ the expression derived in (Aalo, 1995) for the sum of correlated gamma variables which, for this case, becomes

$$p_2(y) = \frac{\left(\frac{y}{\mu_2}\right)^{N_{pilots}-1} \exp\left(-\frac{y}{(1-\sqrt{\rho})\mu_2}\right) {}_1F_1\left(1, N_{pilots}; \frac{\sqrt{\rho} N_{pilots} y}{(1-\sqrt{\rho})(1-\sqrt{\rho} + \sqrt{\rho} N_{pilots})\mu_2}\right)}{(N_{pilots}-1)! (1-\sqrt{\rho})^{N_{pilots}-1} (1-\sqrt{\rho} + \sqrt{\rho} N_{pilots})\mu_2} u(y) \quad (19)$$

where ρ is the correlation coefficient between different received samples which is constant and is defined as

$$\rho = \rho_{(k,l),(k',l')} = \frac{\text{Cov}\left(\left|R_{k,l}^1\right|^2, \left|R_{k',l'}^1\right|^2\right)}{\sqrt{\text{Var}\left(\left|R_{k,l}^1\right|^2\right)\text{Var}\left(\left|R_{k',l'}^1\right|^2\right)}}, \quad (k,l) \neq (k',l') \quad (20)$$

with

$$\text{Cov}\left(\left|R_{k,l}^1\right|^2, \left|R_{k',l'}^1\right|^2\right) = 2\left|S_{k,l}^{p,pilot}\right|^4 \left(E\left[\left|H_{k,l}^{p,1}\right|^2\right]\right)^2 + 2\left|S_{k,l}^{p,pilot}\right|^2 E\left[\left|H_{k,l}^{p,1}\right|^2\right] N_0 + N_0^2 - \mu_2^2 \quad (21)$$

and

$$\text{Var}\left(\left|R_{k,l}^1\right|^2\right) = 2\left|S_{k,l}^{p,pilot}\right|^4 \left(E\left[\left|H_{k,l}^{p,1}\right|^2\right]\right)^2 + 4\left|S_{k,l}^{p,pilot}\right|^2 E\left[\left|H_{k,l}^{p,1}\right|^2\right] N_0 + 2N_0^2 - \mu_2^2 \quad (22)$$

Alternatively, from (Alouini et al., 2001), we can also represent (19) as a single gamma-series

$$p_2(y) = \frac{\lambda_1}{\lambda_{N_{pilots}}} \sum_{t=0}^{\infty} \frac{\delta_t y^{N_{pilots}+t-1} \exp\left(-\frac{y}{\lambda_1}\right)}{\lambda_1^{N_{pilots}+t} (N_{pilots}+t-1)!} \quad (23)$$

with

$$\delta_t = \begin{cases} 1, & t=0 \\ \frac{1}{t} \sum_{i=1}^t \left(1 - \frac{\lambda_1}{\lambda_{N_{pilots}}}\right)^i \delta_{t-i}, & t>0 \end{cases} \quad (24)$$

and

$$\lambda_1 = \mu_2(1 - \sqrt{\rho}); \quad \lambda_{N_{pilots}} = \mu_2 \left[1 + \sqrt{\rho}(N_{pilots} - 1)\right] \quad (25)$$

${}_1F_1(\cdot, \cdot; \cdot)$ is the confluent hypergeometric function (Milton & Stegun, 1964). The weighted intersection of PDFs (16) and (19) or (23) (threshold y^{th}) can be easily found numerically.

5. Medium access control

To evaluate the detection technique presented above we will use the analysis presented in (3GPP TR101 102 v3.2.0, 1998) for the network-assisted diversity multiple access (NDMA) MAC protocol. It is assumed that the users transmit packets to a BS, which is responsible for running most of the calculations and to handle transmission collisions. The BS detects collisions and uses a broadcast control channel to send a collision signal, requesting the users to resend the collided packets the required number of times ($p-1$ for a collision of p packets). The remaining section studies how the throughput is influenced by the block/packet error rate (BLER), and compares the results with the performance of a contention-free scenario, based on TDMA.

5.1 Throughput analysis

Following the NDMA throughput analysis of (3GPP TR101 102 v3.2.0, 1998), we consider a sequence of epochs where epoch is an empty slot or a set of slots where users send the same packet due to a BS request. Denoting P_e as the probability of a user's buffer being empty at the beginning of an epoch, the binomial expressions for the probability of the epoch length for J users are

$$P_{busy}(p) = \binom{J}{p} (1 - P_e)^p P_e^{J-p}, p = 1, 2, \dots, J \quad (26)$$

for a busy epoch and

$$P_{idle}(p) = \begin{cases} P_e^J, & p = 1 \\ 0, & p \neq 1 \end{cases} \quad (27)$$

for an idle epoch. The probability of having a useful epoch is

$$P_{usefull}(p) = \binom{J}{p} (1 - P_e)^p P_e^{J-p} P_D(p)^p \quad (28)$$

where $P_D(p)$ is the frame's correct detection probability (equal to $1 - BLER$) when p users are transmitting. We assume that no detection errors occur in the determination of the number of senders colliding. Finally, the throughput can be defined as

$$R_{NDMA} = \frac{\text{average length of useful epoch}}{\text{average length of busy or idle epoch}} \quad (29)$$

By using (26) and (29), and after some simplifications, we can write

$$R_{NDMA} = \frac{\sum_{p=1}^J p \binom{J-1}{p-1} (1 - P_e)^p P_e^{J-p} P_D(p)^p}{J(1 - P_e) + P_e^J} \quad (30)$$

5.2 Queue analysis

If there are no detection errors at the receiver (i.e., the BS), then the busy and idle epochs have the distributions described by

$$P_{busy}(p) = \binom{J-1}{p-1} (1 - P_e)^{J-1} P_e^{J-p}, 1 \leq p \leq J \quad (31)$$

and

$$P_{idle}(p) = \begin{cases} P_e^{J-1} + (J-1)(1 - P_e)P_e^{J-2}, & p = 1 \\ \binom{J-1}{p} (1 - P_e)^{p-1} P_e^{J-p-1}, & 1 \leq p \leq J-1 \end{cases} \quad (32)$$

where P_e is the unique solution on $[0, 1]$ of the equation (see (3GPP TR101 102 v3.2.0, 1998))

$$\lambda P_e^J + (1 - \lambda J)P_e - (1 - \lambda J) = 0. \quad (33)$$

5.3 Delay analysis

For an M/G/1 queue with vacation the average system delay for a data packet can be expressed as

$$D = \overline{h_{busy}} + \frac{\overline{\lambda h_{busy}^2}}{2(1 - \lambda \overline{h_{busy}})} + \frac{\overline{\lambda h_{idle}^2}}{2\overline{h_{idle}}}, \quad (34)$$

where $\overline{h_{busy}}$, $\overline{h_{busy}^2}$, $\overline{h_{idle}}$ and $\overline{h_{idle}^2}$ are the first and second moments of the busy and idle epoch respectively.

5.4 Comparison with ideal TDMA protocols

Traditional MAC protocols loose packets involved in collisions. The best performance with traditional MAC protocols is achieved when collisions are avoided, with a TDMA (time division multiple access) approach. The throughput for an ideal TDMA protocol depends linearly with the total offered load, and with the probability of correct detection of a single sender, i.e.,

$$R_{TDMA} = \lambda J P_D(1) \quad (35)$$

For large SNR $P_D \approx 1$ and (30) can be written as

$$R_{NDMA} = \frac{J(1 - P_e)}{J(1 - P_e) + P_e^J}. \quad (36)$$

It can be shown that (36) is equal to R_{TDMA} when a Poisson source is used (see (3GPP TR101 102 v3.2.0, 1998)). Therefore, NDMA and TDMA throughputs are the same when no detection errors occur, and converge to one near saturation. However, NDMA outperforms TDMA for low signal to noise ratio values, due to the detection gain for multiple transmissions.

6. Numerical results

In this section we present several performance results concerning multipacket detection for OFDM-based systems. The channel impulse response is characterized by the PDP (Power Delay Profile) based on the Vehicular A environment (3GPP TR101 102 v3.2.0, 1998), although similar results would be obtained for other severely time-dispersive channels. Rayleigh fading was admitted for the different paths. The number of subcarriers employed was $N=256$ with a spacing of 15 kHz and each carrying a QPSK data symbol. The channel encoder was a rate-1/2 turbo code based on two identical recursive convolutional codes characterized by $G(D) = [1 (1+D^2+D^3)/(1+D+D^3)]$. A random interleaver was employed within the turbo encoder. The coded bits were also interleaved before being mapped into a

QPSK constellation. Each information stream was encoded with a block size of 3836 bits which, combined with a pilot insertion spacing of $\Delta N_F = N_{p,\max}$ and $\Delta N_T = 16$ results in a frame composed of 16 OFDM blocks. The power level of the pilots symbols was chosen as

$$E\left[|S_{k,l}^{p,Pilot}|^2\right] / E\left[|S_{k,l}^p|^2\right] = 10 \cdot \log_{10}(N_{p,\max})$$
 ($E[\cdot]$ represents the expected value computed over all positions (k,l) containing pilot symbols in the case of the numerator and over all positions containing data symbols in the case of the denominator) so that the percentage of overall transmitted power spent on the pilots was always the same, independently of the maximum number of users, $N_{p,\max}$.

Regarding the channels for the N_p retransmissions of a given packet, we considered three scenarios: uncorrelated channels (UC), fixed channels (FC), corresponding to a stationary scenario, and variable channels (VC), where the mobile speeds are $v=30\text{km/h}$. Unless otherwise stated, uncorrelated symbol interleavers are assumed for different retransmissions. The performance results are expressed as a function of the E_b/N_0 , where E_b is the average bit energy per packet and N_0 the one-sided power spectral density of the channel noise.

As explained previously, ideally the multipacket separation should be made using MLSO, but an MMSE-based separation is much less complex, especially for larger constellations and/or when a large number of packets collide. These multipacket separation techniques can be combined with IC in an iterative receiver as explained in Section 4.

In Fig. 5 we present the BLER for MMSE and MLSO packet separation schemes with or without IC in the case of a collision involving 2 packets. For the schemes without IC we assumed that there are 12 iteration of the turbo decoder. For the cases with IC we have an initial MLSO/MMSE packet separation step and 3 IC steps, each one with 3 iterations applied inside the channel decoder. Regarding the retransmissions, we assumed the VC scenario (similar conclusions could be drawn for other scenarios). From this figure, it is clear that the best performances are attained when we combine MMSE or MLSO separation with IC (in fact, similar performances are achievable when MMSE or MLSO separation are combined with IC); if we do not employ IC an initial MLSO packet separation allows much better performance than MMSE packet separation). In the following results we will always assume an MMSE packet separation combined with 3 IC steps.

The figure (Fig 6) shows the impact of the symbol interleaving for different numbers of colliding packets. Four retransmission scenarios (VC without interleaving, FC and VC employing different symbol interleavers, and UC) are considered. Regarding the two VC scenarios, it is clear that, for the adopted mobile speeds (30km/h) the channel correlations are too high to allow efficient packet separation if we do not employ different symbol interleavers for different retransmissions. Comparing all the different scenarios, as expected, the performances are better for UC scenarios and worse for FC scenarios, with VC having performances in-between. It is important to highlight the fact that although the FC scenario corresponds to a channel that remains fixed for the retransmissions we can still achieve reliable detection with our receiver due to use of the symbols interleavers in the retransmissions.

In Fig. 7 we show the BLER performance for different values of N_p assuming VC scenario. Clearly, our receiver allows an efficient packet separation. From this figure we can observe performance improvement as we increase N_p , which is a consequence of having adopted the E_b for each packet (the total energy used to transmit a packet is $N_p E_b$, since the total number of versions that were transmitted is N_p).

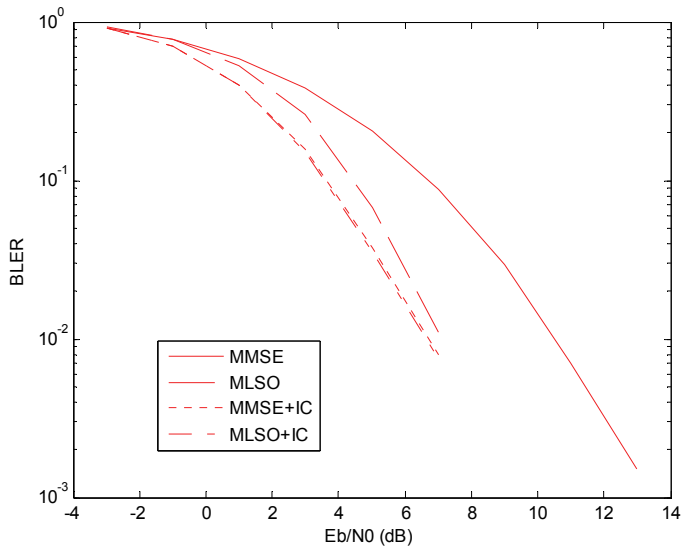


Fig. 5. BLER performance for different packet separation techniques, when $N_p=2$ for VC.

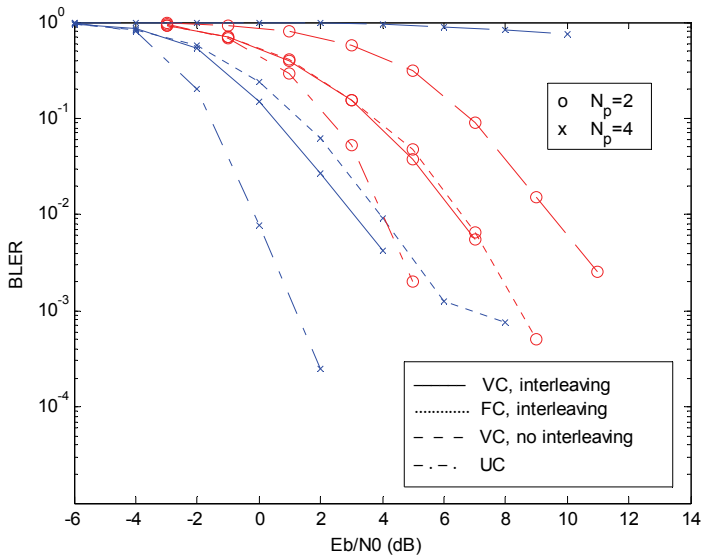


Fig. 6. Impact of using different interleavings for different retransmissions.

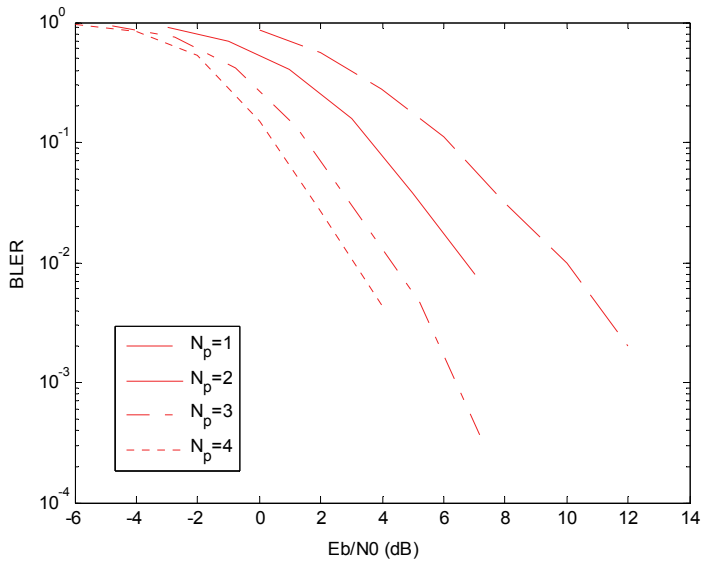


Fig. 7. BLER with different values of N_p for VC.

Using the approach described in Section !!! we present the results regarding the Detection Error Rate (DER) for $N_{p,max}=4$ and $P_e=0.2$ (a high probability of transmission for each user) in Fig. 8.

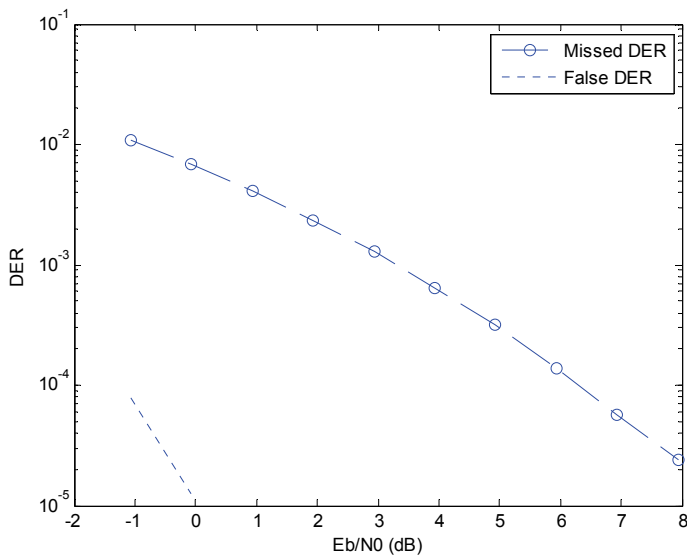


Fig. 8. Detection Error Rate for $N_{p,max}=4$ and $P_e=0.2$.

Curves representing the false DER (false detection of users) and missed DER (users not detected) are shown. It is visible that for $E_b/N_0=2\text{dB}$ that the DER is mostly caused by undetected users (the receiver cannot distinguish them from noise) with an error rate between 0.2-0.3% while false alarms are virtually inexistent.

Next we compare NDMA and TDMA throughputs for the scenario simulated previously. Throughput is calculated as described in Section 5, using BLER obtained above (it should be emphasized that our throughput model does not take into account invalid detection of the number of senders on a collision).

In Fig. 9, Fig. 10 and Fig. 11 show how R_{NDMA} and R_{TDMA} depend on the offered load, for E_b/N_0 values of 2dB, 4dB and 6dB, respectively. The offered load (λ) varies from very light load (10%) until the saturation value (100%), where all bandwidth is required to satisfy the offered load. Results show that NDMA clearly outperforms TDMA for the conditions tested, especially for loads above 60%, with higher differences for lower E_b/N_0 . The reason for this behavior is that our receiver can take full advantage of the overall energy spent to transmit the packet (i.e., the energy for all retransmission attempts). Therefore, the performance of transmitting with success a given packet when we have a collision of several packets is higher than without collisions (as in the TDMA case) due to the BLER performance improvement with larger N_p (as shown in Fig. 7). The only case where our technique is worse than conventional TDMA schemes is for slow-varying channels without symbol interleaving, especially for large system loads, since the correlation between different retransmissions can be very high, precluding an efficient packet separation.

The throughput obtained in fixed or variable channels combined with interleaving is only slightly worse than that obtained in uncorrelated channels.

We would like to point out that although the throughput for high system load can be close to 100%, the corresponding packet delay grows fast for large system loads, since the number of retransmission increases with the number of collisions, and the number of collisions is higher for higher system loads.

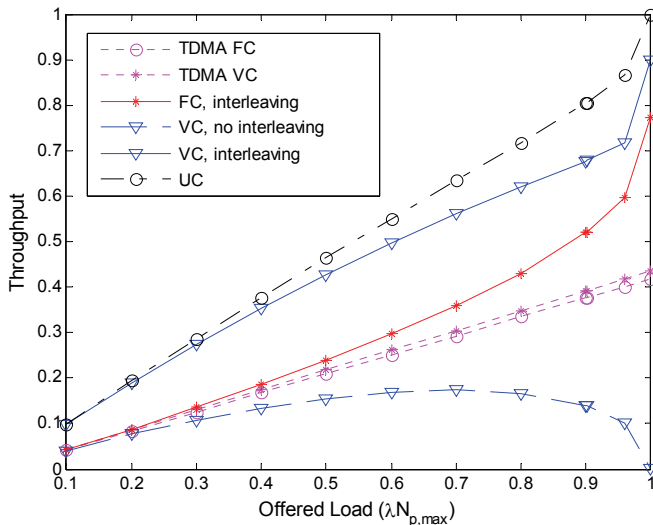


Fig. 9. Throughput when $E_b/N_0=2\text{dB}$.

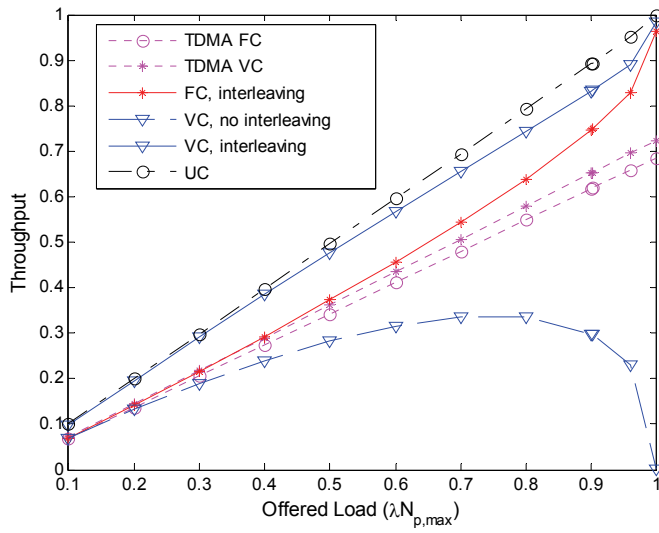


Fig. 10. Throughput when $E_b/N_0=4\text{dB}$.

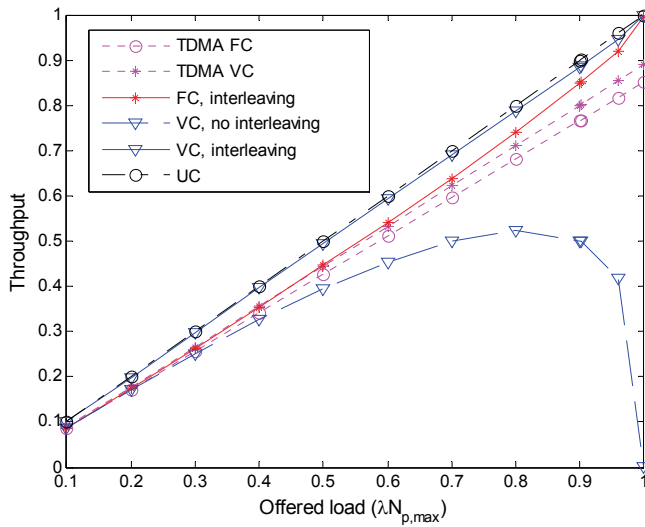


Fig. 11. Throughput when $E_b/N_0=6\text{dB}$.

7. Conclusions

In this chapter we considered a multipacket detection technique to cope with MAC collisions in OFDM-based systems. This technique allows high throughputs, since the total number of transmissions can be equal to the number of packets involved in the collision. Since our packet separation technique requires different channels for different retransmissions we proposed the use of different interleavers for different retransmissions. This allows good performances even slow-varying channels. In fact, we can an efficient packet separation even when the channel remains fixed for all retransmissions. We also included a method to estimate the number of users involved in a collision, as well as the corresponding channel characteristics.

8. Acknowledgments

This work was partially supported by the FCT - Fundação para a Ciência e Tecnologia (pluriannual funding and U-BOAT project PTDC/EEA-TEL/67066/2006), and the C-MOBILE project IST-2005-27423.

9. References

- 3GPP. 25.212-v6.2.0 (2004) "Multiplexing and Channel Coding (FDD),2004 , 3rd Generation Partnership Project, Sophia- Antipolis, France.
- 3GPP. TR25.814 (2006). "Physical Layers Aspects for Evolved UTRA" ,2006 ,3rd Generation Partnership Project, Sophia- Antipolis, France.
- 3GPP TR101 112 v3.2.0 (1998). 'Selection procedures for the choice of radio transmission technologies of UMTS'. 1998 , 3rd Generation Partnership Project, Sophia-Antipolis, France.
- Aalo, V. A. (1995). "Performance of maximal-ratio diversity systems in a correlated Nakagami-fading environment". *IEEE Trans. Commun.*, vol. 43, pp. 2360-2369, 0090-6778.
- Abramowitz, M., & Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 0-486-61272-4, New York, USA
- Alouini, M. -S.; Abdi, A. & Kaveh, M. (2001). "Sum of Gamma Variates and Performance of Wireless Communication Systems Over Nakagami-Fading Channels". *IEEE Trans. On Veh. Tech.*, pp. vol. 50, no. 6, pp.1471-1480, 2001, 1751-8628.
- Cavers, J. K. (1991). "An analysis of Pilot Symbol Assisted Modulation for Rayleigh Fading Channels". *IEEE Trans. On Veh. Tech.*, vol. 40, no. 4, pp.686-693, November, 1991, 0018-9545.
- Cimini, L. (1985). "Analysis and Simulation of a Digital Mobile Channel using Orthogonal Frequency Division Multiplexing". *IEEE Trans. on Comm*, Vol. 33, No. 7, pp.665-675 ,July , 1985.
- Dinis, R.; Carvalho; P., Bernardo; L., Oliveira; R., Serrazina, M. & Pinto, P. (2007). " Frequency-Domain Multipacket Detection: A High Throughput Technique for SC-FDE Systems ". *IEEE GLOBECOM'07.*, pp.4619-4624, 978-1-4244-1043-9, Washington DC., USA, December 2007.

- Dinis, R.; Serrazina, M., & Carvalho, P. (2007). "An Efficient Detection Technique for SC-FDE Systems with Multiple Packet Collisions", *IEEE ICCCN'07*, pp.402-407 , ,Turtle Bay, USA, September 2007.
- Kay, M. S. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, 0-13-3457117, Englewood Cliffs.
- Kim, Y. -S.; Kim, C. -J.; Jeong, G. -Y.; Bang, Y. -S.; Park, H. -K. & Choi, S. S. (1997). "New Rayleigh fading channel estimator based on PSAM channel sounding technique". *IEEE International Conf. on Comm.* ,Vol. 3 , pp.1518-1520., 0-7803-3925-8 Montreal, Canada , June 1997.
- Nee, R. van & Prasad, R. (2000). "*OFDM for Wireless Multimedia Communications*". Artech House, 978-0890065303, Norwood, MA, USA.
- Souto, N.; Correia, A.; Dinis, R.; Silva, J. C. & Abreu, L. (2008). "Multiresolution MBMS transmissions for MIMO UTRA LTE systems". *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting.*, pp.1-6, 978-1-4244-1648-6, Las Vegas, USA, June 2008.
- Tsatsanis, M.; Zhang, R. & Banerjee, S. (2000). Network-Assisted Diversity for Random Access Wireless Networks, *IEEE* ,48, 3,Mar. 2000, pp.702-708, 1053-587X
- Valenti, M. C. (2001). "Iterative Channel Estimation and Decoding of Pilot Symbol Assisted Turbo Codes Over Flat-Fading Channels". *IEEE Journal on Selected Areas in Communications*,Vol. 19 ,No. 9 ,September 2001 , pp. 1697-1705.
- Wang, X.; Yu, Y. & Giannakis, G. (2005). "A Robust High-Throughput Three Algorithm Using Successive Interference Cancellation". *IEEE GLOBECOM'05.*,Vol.6, pp. - 3601, St. Louis, USA, 0-7803-9414-3.
- Yu, Y. & Giannakis, G. (2005). "SICTA: A 0.693 Contention Tree Algorithm Using Successive Interference Cancellation". *IEEE INFOCOM'05.* ,Vol. 3 , 1908 - 1916, March, 2005, 0743-166X
- Zhang, R. & Tsatsanis, M. (2002). "Network-Assisted Diversity Multiple Access in Time-Dispersive Channels". *IEEE Trans. On Comm.*, Vol. 50, No. 4, pp. 623-632.,April 2002, 0090-6778

Joint Subcarrier Matching and Power Allocation for OFDM Multihop System

Wenyi Wang and Renbiao Wu

*Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China
China*

1. Introduction

Relay networks have recently attracted extensive attention due to its potential to increase coverage area and channel capacity. In a relay network, a source node communicates with a destination node with the help of the relay node. The performances of improving the channel capacity and coverage area have been explored and evaluated in the literature (Sendonaris et al., 2003)-(Laneman et al., 2004). There are two main forwarding strategies for relay node: amplify-and-forward (AF) and decode-and-forward (DF) (Laneman et al., 2004). The AF cooperative relay scheme was developed and analyzed in (Shastri & Adve, 2005), where a significant gain in the network lifetime due to node cooperation was shown. Power allocation is studied and compared for AF and DF relaying strategies for relay networks, which improves the channel capacity (Serbetli & Yener, 2006). However, DF means that the signal is decoded at the relay and recoded for retransmission. It is different from AF, where the signal is magnified to satisfy the power constraint and forwarded at the relay. This has the main advantage that the transmission can be optimized for different links, separately. In this chapter, the relay strategy DF is used.

In wideband systems, orthogonal frequency division multiplexing (OFDM) is a mature technique to mitigate the problems of frequency selectivity and intersymbol interference. The optimization of power allocation for different subcarriers offers substantial gain to the system performance. Therefore, the combination of relay network and OFDM modulation is an even more promising way to improve capacity and coverage area. However, as the fading gains for different channels are mutually independent, the subcarriers which experience deep fading over the source-relay channel may not be in deep fading over the relay-destination channel. This motivates us to consider adaptive subcarrier matching and power allocation schemes, where the bits on the subcarriers from the source to the relay are reassigned to the subcarriers from the relay to the destination. The system architecture of OFDM two-hop relay system is demonstrated in the Fig.1.

A fundamental analysis of cooperative relay systems was done by Kramer (Kramer et al., 2006), who has given channel capacity of several schemes. Relaying for OFDM systems was considered theoretically in (Shastri & Adve, 2005). Multi-user OFDM relay networks were studied by Zhu (Zhu et al., 2005), where the subcarrier was allocated to transmit own information and forward other nodes' information. Relay selection in OFDM relay networks was studied by Dai (Dai et al., 2007), which indicated the maximum diversity by selecting different relay for the different subcarrier. Radio resource allocation algorithm for relay

The worse subcarrier decreases the capacity of the matched subcarriers without subcarrier matching.

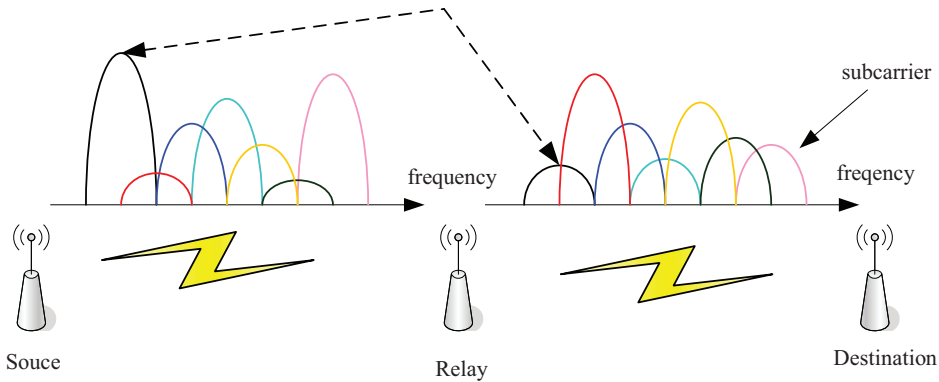


Fig. 1. System architecture of OFDM two-hop relay system

aided cellular OFDMA system was done in (Kaneko & Popovski, 2007). Adaptive relaying scheme for OFDM that taking channel state information into account has been proposed in (Herdin, 2006), where subcarrier matching was considered for OFDM amplify-and-forward scheme but the power allocation was not considered. Performances of OFDM dual-hop system with and without subcarrier matching were studied in (Suraweera & Armstrong, 2007) and (Athaudage et al., 2008), separately. The problems of resource allocation were considered in OFDMA cellular and OFDMA multihop system (Pischella & Belfiore, 2008) and (Kim et al., 2008). Bit loading algorithms were studied in (Ma et al., 2008) and (Gui et al., 2008). The subcarrier matching was also utilized to improve capacity in cognitive radio system (Pandharipande & Ho, 2007)-(Pandharipande & Ho, 2008).

In this chapter, the resource allocation problem is studied to maximize the system capacity by joint subcarrier matching and power allocation for the system with system-wide and separate power constraints. The schemes of optimal joint subcarrier matching and power allocation are proposed. All the proposed schemes perform better than the several other schemes, where there is no subcarrier matching or no power allocation.

The rest of this chapter is organized as follows. Section 2 discusses the optimal subcarrier matching and power allocation for the system with system-wide power constraint. Section 3 discusses the optimal subcarrier matching and power allocation for the system with separate power constraints. Section 6 compares the capacities of optimal schemes with that of several other schemes. Conclusions are drawn in section 5.

2. The system with system-wide power constraint

2.1 System architecture and problem formulation

An OFDM multihop system is considered where the source communicates with the destination using a single relay. The relay strategy is decode-and-forward. All nodes hold one antenna. It is assumed that the destination receives signal only from the relay but not from the source because of distance or obstacle. A two-stage transmission protocol is adopted. This means that the communication between the source and the destination covers two equal time slots. Fig.2 shows the block diagram of joint subcarrier matching and power allocation. The source transmits an OFDM symbol over the source-relay channel during the first time slot. At the same time, the relay receives and decodes the symbol. During the

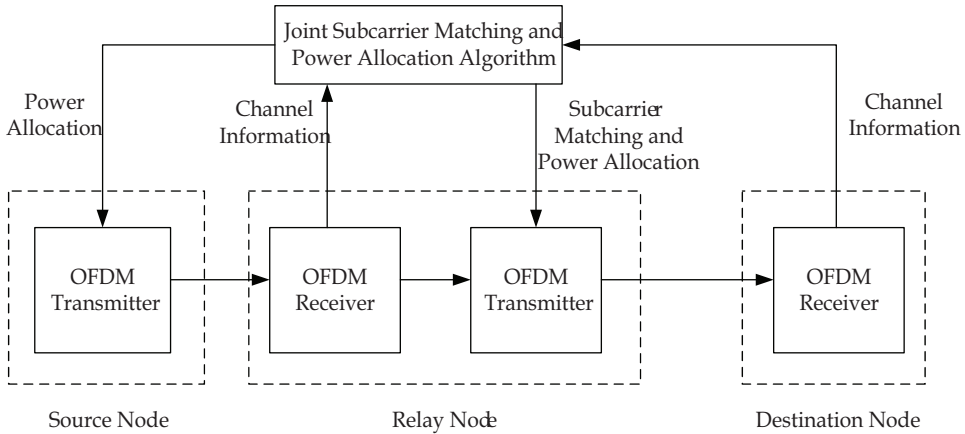


Fig. 2. Block diagram of joint subcarrier matching and power allocation

second time slot, the relay reencodes the signal with the same codebook as the one used at the source, and transmits it towards the destination over the relay-destination channel. The destination decodes the signal based on the received signal only from the relay. Furthermore, full channel state information (CSI) is assumed. The source transmits the signal to the relay with power allocation among the subcarriers based on the algorithm of joint subcarrier matching and power allocation. The relay receives the signal and decodes the signal. Then, the relay reorders the subcarrier to match subcarrier, and allocates power among the subcarriers according to the algorithm of joint subcarrier matching and power allocation. At last, the destination decodes the signal.

In this chapter, it is assumed that the different channels experience independent fading. The system consists of N subcarriers with total system power constraint. The power spectral densities of additive white Gaussian noise (AWGN) are equal at the relay and the destination. The channel capacity of the subcarrier i over the source-relay channel is given as follows

$$R_{s,i}(P_{s,i}) = \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) \quad (1)$$

where $P_{s,i}$ is the power allocated to the subcarrier i ($1 \leq i \leq N$) at the source, $h_{s,i}$ is the corresponding channel power gain, and N_0 is the power spectral density of AWGN. Similarly, the channel capacity of the subcarrier j over the relay-destination channel is given as follows

$$R_{r,j}(P_{r,j}) = \frac{1}{2} \log_2 \left(1 + \frac{P_{r,j} h_{r,j}}{N_0} \right) \quad (2)$$

where $P_{r,j}$ is the power allocated to the subcarrier j ($1 \leq j \leq N$) at the relay, and $h_{r,j}$ is the corresponding channel power gain.

Consequently, when the subcarrier i over the source-relay channel is matched to the subcarrier j over the relay-destination channel, the channel capacity of this subcarrier pair is given as follows

$$R_{ij} = \min\{R_{s,i}(P_{s,i}), R_{r,j}(P_{r,j})\} \quad (3)$$

Theoretically, the bits transmitted at the source can be reallocated to the subcarriers at the relay in arbitrary way. But for simplification, an additional constraint is that the bits transported on a subcarrier over the source-relay channel can be reallocated to only one subcarrier over the relay-destination channel, i.e., only one-to-one subcarrier matching is permitted. This means that the bits on different subcarriers over the source-relay channel will not be reallocated to the same subcarrier at the relay.

For the optimal joint subcarrier matching and power allocation problem, we can formulate it as an optimization problem. The optimization problem is given as

$$\begin{aligned} & \max_{P_{s,i}, P_{r,j}, \rho_{ij}} \sum_{i=1}^N \min \left\{ R_{s,i}(P_{s,i}), \sum_{j=1}^N \rho_{ij} R_{r,j}(P_{r,j}) \right\} \\ & \text{subject to } \sum_{i=1}^N P_{s,i} + \sum_{j=1}^N P_{r,j} \leq P_{tot} \\ & P_{s,i}, P_{r,j} \geq 0, \forall i, j \\ & \sum_{j=1}^N \rho_{ij} = 1, \rho_{ij} = \{0, 1\}, \forall i, j \end{aligned}$$

where P_{tot} is the total system power constraint, and ρ_{ij} can only be either 1 or 0, indicating whether the bits transmitted on the subcarrier i at the source are retransmitted on the subcarrier j at the relay. The last constraint shows that only one-to-one subcarrier matching is permitted. By introducing the parameter C_i , the optimization problem can be transformed into

$$\begin{aligned} & \max_{P_{s,i}, P_{r,j}, \rho_{ij}, C_i} \sum_{i=1}^N C_i \\ & \text{subject to } R_{s,i}(P_{s,i}) \geq C_i \\ & \sum_{j=1}^N \rho_{ij} R_{r,j}(P_{r,j}) \geq C_i \\ & \sum_{i=1}^N P_{s,i} + \sum_{j=1}^N P_{r,j} \leq P_{tot} \\ & P_{s,i}, P_{r,j} \geq 0, \forall i, j \\ & \sum_{j=1}^N \rho_{ij} = 1, \rho_{ij} = \{0, 1\}, \forall i, j \end{aligned}$$

Consequently the original maximization problem is transformed into a mixed binary integer programming problem. It is prohibitive to find the global optimum in terms of computational complexity. However, when ρ_{ij} is given, the objective function and all constraint functions are convex, so the optimization problem is a convex optimization problem. Then the optimal power allocation can be achieved by interior-point algorithm. Therefore, the optimal joint subcarrier matching and power allocation can be found by finding the largest objective function among all subcarrier matching possibilities, and the corresponding subcarrier matching as well as power allocation is jointly optimal. But, it has

been proved to be NP-hard and is fundamentally difficult (Korte & Vygen, 2002). In next subsection, with analytical argument, a low complexity and optimal joint subcarrier matching and power allocation scheme is given, where the optimal subcarrier matching is to match subcarriers by the order of the channel power gains and the optimal power allocation among the subcarrier pairs is based on water-filling.

2.2 Optimal joint subcarrier matching and power allocation for the system including two subcarriers

Supposing that the system includes only two subcarriers ($N = 2$): the channel power gains over the source-relay channel are $h_{s,1}$ and $h_{s,2}$, and the channel power gains over the relay-destination channel are $h_{r,1}$ and $h_{r,2}$. Without loss of generality, we assume that $h_{s,1} \leq h_{s,2}$ and $h_{r,1} \leq h_{r,2}$. The total system power constraint is also P_{tot} . As discussed in the subsection 2.1, the optimal joint subcarrier matching and power allocation can be found by two steps: (1) for every matching possibility (i.e., ρ_j is given), find the optimal power allocation and the total channel capacity; (2) compare the all the total channel capacities, the largest one is the largest total channel capacity, whose subcarrier matching and power allocation are jointly optimal. But this process is prohibitive in terms of complexity. In this subsection, an analytical argument is given to prove that the optimal subcarrier matching is to match subcarrier by the order of the channel power gains and the optimal power allocation between the matched subcarrier pairs is based on water-filling. The more important is that they are jointly optimal.

Before giving the scheme, the equivalent channel power gain is given for any matched subcarrier pair. For any given matched subcarrier pair, with the total power constraint, an equivalent channel power gain can be obtained by the following proposition, whose channel capacity is equivalent to the channel capacity of this subcarrier pair.

Proposition 1: For any given matched subcarrier pair, with total power constraint, an equivalent subcarrier channel power gain (e.g., h'_i) can be obtained, which is related to the channel power gains (e.g., $h_{s,i}$ and $h_{r,j}$) of the subcarrier pair as follows

$$\frac{1}{h'_i} = \frac{1}{h_{s,i}} + \frac{1}{h_{r,j}} \quad (4)$$

Proof: With the total power constraint P'_i , the channel capacity of this subcarrier pair is

$$R'_i = \max_{P_{s,i}} \min \left\{ \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right), \frac{1}{2} \log_2 \left(1 + \frac{(P'_i - P_{s,i}) h_{r,j}}{N_0} \right) \right\} \quad (5)$$

where $P_{s,i}$ is the power allocated to the subcarrier i at the source, $P'_i - P_{s,i}$ is the remainder power allocated to the subcarrier j at the relay.

The first term is a monotonically increasing function of $P_{s,i}$ and the second term is a monotonically decreasing function of $P_{s,i}$. Therefore, the optimal power allocation between the corresponding subcarriers can be obtained easily

$$\frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) = \frac{1}{2} \log_2 \left(1 + \frac{(P'_i - P_{s,i}) h_{r,j}}{N_0} \right) \quad (6)$$

which means that $h_{s,i}P_{s,i} = h_{r,j}(P'_i - P_{s,i})$. As a result, the channel capacity of the subcarrier pair is

$$R'_i = \frac{1}{2} \log_2 \left(1 + \frac{h_{s,i}h_{r,j}P'_i}{(h_{s,i} + h_{r,j})N_0} \right) \quad (7)$$

It can be seen that the subcarrier pair is equivalent to a single subcarrier channel with the same total power constraint. The equivalent channel power gain h'_i can be expressed

$$h'_i = \frac{h_{s,i}h_{r,j}}{h_{s,i} + h_{r,j}} \quad (8)$$

or

$$\frac{1}{h'_i} = \frac{1}{h_{s,i}} + \frac{1}{h_{r,j}} \quad (9)$$

Here, there are two ways to match the subcarriers: (i) the subcarrier 1 over the source-relay channel is matched to the subcarrier 1 over the relay-destination channel, and the subcarrier 2 over the source-relay channel is matched to the subcarrier 2 over the relay-destination channel (i.e., $h_{s,1} \sim h_{r,1}$ and $h_{s,2} \sim h_{r,2}$); (ii) the subcarrier 1 over the source-relay channel is matched to the subcarrier 2 over the relay-destination channel, and the subcarrier 2 over the source-relay channel is matched to the subcarrier 1 over the relay-destination channel (i.e., $h_{s,1} \sim h_{r,2}$ and $h_{s,2} \sim h_{r,1}$).

For the two ways of matching subcarriers, the equivalent channel power gains are denoted as $h'_{k,i}$ which can be obtained easily based on the proposition 1. Here, k implies the method of matching subcarrier and i is the equivalent subcarrier index. Then, the power allocation between the subcarrier pairs can be reformed as follow

$$\begin{aligned} & \max_{P'_i} \quad \sum_{i=1}^2 \frac{1}{2} \log_2 \left(1 + \frac{h'_{k,i}P'_i}{N_0} \right) \\ & \text{subject to} \quad \sum_{i=1}^2 P'_i \leq P_{tot} \end{aligned}$$

where P'_i is the power allocated to the equivalent subcarrier i .

It's clear that the optimal power allocation is based on water-filling (Cover & Thomas, 1991). Therefore, once the subcarrier matching is provided, the optimal power allocation is easily obtained. The remainder task is to decide which way of subcarrier matching is better. The better method can be found by getting the channel capacities of the two ways and comparing them. But, here, we give an analytical argument to prove that the optimal subcarrier matching way is the first way.

Before giving the optimal subcarrier matching way, based on the proposition 1, we can get following lemma.

Lemma 1: For the two ways of matching subcarrier, the relationship between the equivalent channel power gains can be expressed

$$\frac{1}{h'_{1,1}} + \frac{1}{h'_{1,2}} = \frac{1}{h'_{2,1}} + \frac{1}{h'_{2,2}} \quad (10)$$

Proof: Based on the proposition 1, the equivalent channel power gains of the two ways can be expressed $\frac{1}{h'_{1,1}} = \frac{1}{h_{s,1}} + \frac{1}{h_{r,1}}$, $\frac{1}{h'_{1,2}} = \frac{1}{h_{s,2}} + \frac{1}{h_{r,2}}$ and $\frac{1}{h'_{2,1}} = \frac{1}{h_{s,1}} + \frac{1}{h_{r,2}}$, $\frac{1}{h'_{2,2}} = \frac{1}{h_{s,2}} + \frac{1}{h_{r,1}}$. By summing up the corresponding terms, it is clearly that the relationship can be derived. By making use of the lemma 1, the following proposition can be proved, which states the optimal subcarrier matching way.

Proposition 2: For the system including two subcarriers, the optimal subcarrier matching is to match the subcarriers by the order of channel power gains. Together with the optimal power allocation for this subcarrier matching, they are optimal joint subcarrier matching and power allocation. In this case, the optimal subcarrier matching is as $h_{s,1} \sim h_{r,1}$ and $h_{s,2} \sim h_{r,2}$.

Proof: For the two ways of matching subcarrier, based on the lemma 1, the equivalent channel power gains satisfy the following constraint, $\frac{1}{h'_{k,1}} + \frac{1}{h'_{k,2}} = H (H \geq 0)$, where the parameter H is a constant. For the the first way, we can get $\frac{1}{h'_{1,1}} - \frac{1}{h'_{1,2}} = x_1 (H \geq x_1 \geq 0)$. For the second way, without loss of generality, it is assumed that $\frac{1}{h'_{2,1}} - \frac{1}{h'_{2,2}} = x_2 (H \geq x_2 \geq 0)$.

Therefore, the $h'_{k,i}$ can be expressed as $h'_{k,1} = \frac{2}{H+x_k}$ and $h'_{k,2} = \frac{2}{H-x_k}$. The corresponding total channel capacity is

$$R_{tot,k}(P'_1, P'_2) = \frac{1}{2} \log_2 \left(1 + \frac{P'_1}{(H+x_k) \frac{N_0}{2}} \right) + \frac{1}{2} \log_2 \left(1 + \frac{P'_2}{(H-x_k) \frac{N_0}{2}} \right) \quad (11)$$

For denotation simplicity, we denote $\frac{N_0}{2}$ as σ_2^2 . The partial derivative of the channel capacity with respect to x_k can be gotten by making use of $P'_2 = P_{tot} - P'_1$

$$\frac{\partial R_{tot,k}(P'_1, P'_2)}{\partial x_k} = \frac{1}{2 \ln 2} \frac{2P'_1 x_k (P_{tot} - P'_1) + (H^2 \sigma_2^2 - x_k^2 \sigma_2^2) (P_{tot} - 2P'_1) + 2P_{tot} H x_k \sigma_2^2}{(H^2 - x_k^2) [(H+x_k) \sigma_2^2 + P'_1] [(H-x_k) \sigma_2^2 + (P_{tot} - P'_1)]} \quad (12)$$

It is noted that, because of $h'_{k,1} \leq h'_{k,2}$, $P'_1 \leq \frac{1}{2} P_{tot}$. Therefore, it is clear that $\frac{\partial R_{tot,k}(P'_1, P'_2)}{\partial x_k}$ is

greater than 0. Therefore, the total channel capacity is a monotonically increasing function of x_k . This means that, the larger is the difference between the equivalent channel power gains, the larger is the total channel capacity. At the same time, it is clearly that the difference between the equivalent channel power gains of the first way is larger than the one of the second way. Therefore, the relationship of the total channel capacities of the two ways can be expressed

$$R_{tot,2}(P'_1, P'_2) \leq R_{tot,1}(P'_1, P'_2) \quad (13)$$

Therefore, we can get the following relationship

$$\max_{P'_i} R_{tot,2}(P'_1, P'_2) = R_{tot,2}(\bar{P}'_1, \bar{P}'_2) \leq R_{tot,1}(\bar{P}'_1, \bar{P}'_2) = \max_{P'_i} R_{tot,1}(P'_1, P'_2) \quad (14)$$

where \bar{P}'_1 and \bar{P}'_2 are the optimal power allocation for the first term. Note that the first term is the total channel capacity of the first way and the last term is the one of the second way. It proves that the first way, whose difference between the equivalent channel power gains is larger, is optimal subcarrier matching way. The more important is that, as the total channel capacity of the first way is the larger one, this subcarrier matching and the corresponding power allocation are the optimal joint subcarrier matching and power allocation. Specially, the optimal subcarrier matching is to match subcarriers by the order of the channel power gains. The optimal joint subcarrier matching and power allocation scheme has been given by now. Specially, the optimal subcarrier matching is to match the subcarriers by the order of the channel power gains and the optimal power allocation between the matched subcarrier pairs is according to the water-filling. The power allocation between the matched subcarrier pair is to make the channel capacities of the two subcarriers equivalent.

2.3 Optimal joint subcarrier matching and power allocation for the system including unlimited number of subcarriers

This subsection extends the method in the subsection 2.2 to the system including unlimited number of the subcarriers. The number of the subcarriers is finite, where the subcarrier channel power gains are $h_{s,i}(i \geq 2)$ and $h_{r,j}(j \geq 2)$. First, the optimal power allocation among the matched subcarrier pair is proposed for given subcarrier matching. Second, we prove that the subcarrier matching by the order of the channel power gains is optimal.

When the subcarrier matching is given, the equivalent channel gains of the subcarrier pairs can be gotten based on the proposition 1, e.g., $h'_i(1 \leq i \leq N)$. The power allocation can be formulated as

$$\begin{aligned} \max_{P'_i} \quad & \sum_{i=1}^N \frac{1}{2} \log_2 \left(1 + \frac{h'_i P'_i}{\sigma_N^2} \right) \\ \text{subject to} \quad & \sum_{i=1}^N P'_i \leq P_{tot} \end{aligned} \quad (15)$$

where the $\sigma_N^2 = N_0$. It is clearly that the power allocation is also based on water-filling. Therefore, the optimal power allocation among the matched subcarrier pairs is according to the water-filling.

Here, without loss of generality, the channel power gains are assumed $h_{s,i} \leq h_{s,i+1}$ and $h_{r,j} \leq h_{r,j+1}$. The following proposition gives the optimal subcarrier matching.

Proposition 3: For the system including unlimited number of the subcarriers, the optimal subcarrier matching is

$$h_{s,i} \sim h_{r,i} \quad (16)$$

Together the optimal power allocation for this subcarrier matching, they are optimal joint subcarrier matching and power allocation

Proof: This proposition will be proved in the contrapositive form. Assuming that there is a subcarrier matching method whose matching result including two matched subcarrier pairs $h_{s,i} \sim h_{r,i+n}$ and $h_{s,i+n} \sim h_{r,i}(n > 0)$, which means that $h_{s,i} \leq h_{s,i+n}$, $h_{r,i} \leq h_{r,i+n}$, and the total capacity is larger than that of the matching method in proposition 3.

When the power allocated to other subcarrier pairs and the other subcarrier matching are constant, the total channel capacity of this two subcarrier pair can be improve based on proposition 2, which imply the channel capacity can be improved by rematching the subcarriers to $h_{s,i} \sim h_{r,i}$ and $h_{s,i+n} \sim h_{r,i+n}$. It is contrary to the assumption. Therefore, there is no subcarrier matching way is better than the way in proposition 3. At the same time, as the total capacity of this subcarrier matching and the corresponding optimal power allocation scheme is the largest, this subcarrier matching together with the corresponding optimal power allocation are the optimal joint subcarrier matching and power allocation.

For the system including unlimited number of the subcarriers, the optimal joint subcarrier matching and power allocation scheme has been given by now. Here, the steps are summarized as follow

- Step 1. Sort the subcarriers at the source and the relay in ascending order by the permutations π and π' , respectively. The process is according to the channel power gains, i.e., $h_{s,\pi(i)} \leq h_{s,\pi(i+1)}$, $h_{r,\pi'(i)} \leq h_{r,\pi'(i+1)}$.
- Step 2. Match the subcarriers into pairs by the order of the channel power gains (i.e., $h_{s,\pi(i)} \sim h_{r,\pi'(i)}$), which means that the bits transported on the subcarrier $\pi(i)$ over the sourcerelay channel will be retransmitted on the subcarrier $\pi'(i)$ over the relay-destination channel.
- Step 3. Based on the proposition 1, get the equivalent channel power gain $h'_{\pi(i)}$ according to the matched subcarrier pair, i.e., $h'_{\pi(i)} = \frac{h_{s,\pi(i)}h_{r,\pi'(i)}}{h_{s,\pi(i)}+h_{r,\pi'(i)}}$.
- Step 4. For the equivalent channel power gains, the power allocation is based on water-filling as follow

$$P'_{\pi(i)} = \left(\frac{1}{2\lambda \ln 2} - \frac{\sigma_N^2}{h'_{\pi(i)}} \right)^+ \quad (17)$$

where $(a)^+ = \max(a,0)$ and λ can be found by the following equation

$$\sum_{i=1}^N P'_{\pi(i)} = P_{tot} \quad (18)$$

The power allocation between the subcarriers in the matched subcarrier pair is as follow

$$P_{s,\pi(i)} = \frac{h_{r,\pi'(i)} P'_{\pi(i)}}{h_{s,\pi(i)} + h_{r,\pi'(i)}} \quad (19)$$

$$P_{r,\pi'(i)} = \frac{h_{s,\pi(i)} P'_{\pi(i)}}{h_{s,\pi(i)} + h_{r,\pi'(i)}} \quad (20)$$

- Step 5. The total system channel capacity is

$$R_{tot} = \frac{1}{2} \sum_{i=1}^N \log_2 \left(1 + \frac{h'_{\pi(i)} P'_{\pi(i)}}{\sigma_N^2} \right) \quad (21)$$

3. The system with separate power constraints

3.1 System architecture and problem formulation

The system architecture adopted in this section is same as the forward section. The difference is the power constraints are separate at the source node and relay node.

It is also noted that there are three ways for the relay to forward the information to the destination. The first is that the relay decodes the information on all subcarriers and reallocates the information among the subcarriers, then forwards the information to the destination. Here, the relay has to reallocate the information among the subcarriers. At the same time, as the number of bits reallocated to a subcarrier are different as that of any subcarrier at the source, different modulation and code type have to be chosen for every subcarrier at the relay. The second is that the information on a subcarrier can be forwarded on only one subcarrier at the relay, but the information on a subcarrier is only forwarded by the same subcarrier. However, as independent fading among subcarriers, it reduces the system capacity. The third is the same as the second according to the information on a subcarrier forwarded on only one subcarrier, but it can be a different subcarrier. Here, for the matched subcarrier pair, as the bits forwarded at the relay are same as that at the source, the relay can utilize the same modulation and code as the source. It means that the bits of different subcarrier may be for different destination. Another example is relay-based downlink OFDMA system. In this system, the second hop consists of multiple destinations where the relay forwards the bits to the destinations based on OFDMA. For this system, subcarrier matching is more preferable than bits reallocation. The bits reallocation at the relay will mix the bits for different destinations. The destination can not distinguish what bits belong to it.

According to the system complexity, the first is the most complex as information reallocation among all subcarriers; the third is more complex than the second as the third has a subcarrier matching process and the second has no it. On the other hand, according to the system capacity, the first is the greatest one without loss by reallocating bits; the third is greater than the second by the subcarrier matching. The capacity of matched subcarrier is restricted by the worse subcarrier because of different fading. In this section, the third way is adopted, whose complexity is slight higher than the second. The subcarrier matching is very simple by permutation, and the system capacity of the third is almost equivalent to the greatest one according to the first and greater than that of the second. The block diagram of system is demonstrated in the Fig.3.

Throughout this section, we assume that the different channels experience independent fading. The system consists of N subcarriers with individual power constraints at the source and the relay, e.g., P_s and P_r . The power spectrum density of additive white Gaussian noise (AWGN) on every subcarrier are equal at the source and the relay.

To provide the criterion for capacity comparison, we give the upper bound of system capacity. Making use of the max-flow min-cut theory (Cover & Thomas, 1991), the upper bound of the channel capacity can be given as

$$C_{upper} = \min \left\{ \max_{P_{s,i}} \sum_{i=1}^N R_{s,i}(P_{s,i}), \max_{P_{r,j}} \sum_{j=1}^N R_{r,j}(P_{r,j}) \right\} \quad (22)$$

It is clear that the optimal power allocations at the source and the relay are according to the water-filling algorithm. By separately performing water-filling algorithm at the source and the relay, the upper bound can be obtained. According to the upper bound, the power allocations are given as following

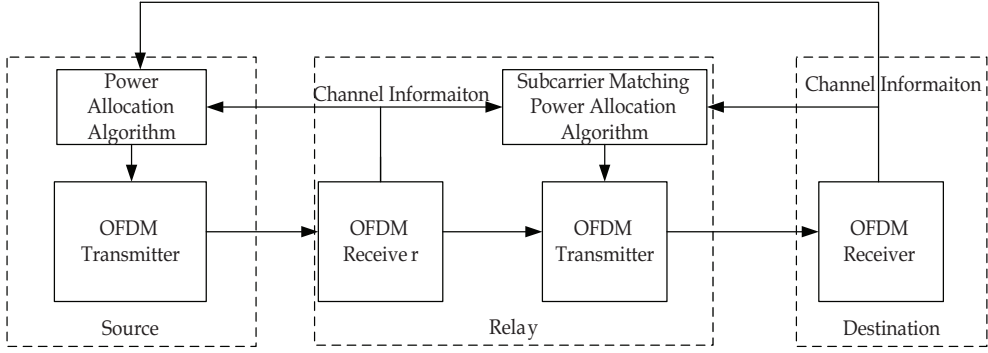


Fig. 3. Details of algorithm block diagram of joint subcarrier matching and power allocation

$$P_{s,i}^{up} = \frac{1}{\lambda_s} - \frac{N_0}{h_{s,i}} \quad (23)$$

$$P_{r,j}^{up} = \frac{1}{\lambda_r} - \frac{N_0}{h_{r,j}} \quad (24)$$

where $P_{s,i}^{up}$ and $P_{r,j}^{up}$ are the power allocations for i and j at the source and the relay. The parameters λ_s and λ_r can be obtained by the following equations

$$\sum_{i=1}^N P_{s,i}^{up} = P_s \quad (25)$$

$$\sum_{j=1}^N P_{r,j}^{up} = P_r \quad (26)$$

Here, the details are omitted, which can be referred to the reference (Cover & Thomas, 1991). Theoretically, the bits transmitted at the source can be reallocated to the subcarriers at the relay in arbitrary way, which is the first way mentioned. However, to simplify system architecture, an additional constraint is that the bits transported on a subcarrier from the source to the relay can be reallocated to only one subcarrier from the relay to the destination, i.e., only one-to-one subcarrier matching is permitted. This means that the bits on different subcarriers at the source will not be forwarded to the same subcarrier at the relay. Later, simulations will show that this constraint is approximately optimal.

The problem of optimal joint subcarrier and power allocation can be formulated as follows

$$\begin{aligned} & \max_{P_{s,i}, P_{r,j}, \rho_{ij}} \sum_{i=1}^N \min \left\{ R_{s,i}(P_{s,i}), \sum_{j=1}^N \rho_{ij} R_{r,j}(P_{r,j}) \right\} \\ & \text{subject to } \sum_{i=1}^N P_{s,i} \leq P_s, \sum_{j=1}^N P_{r,j} \leq P_r \\ & P_{s,i}, P_{r,j} \geq 0, \forall i, j \\ & \sum_{j=1}^N \rho_{ij} = 1, \rho_{ij} = \{0, 1\}, \forall i, j \end{aligned}$$

where ρ_{ij} , being either 1 or 0, is the subcarrier matching parameter, indicating whether the bits transmitted in the subcarrier i at the source are retransmitted on the subcarrier j at the relay. Here, the objective function is system capacity. The first two constraints are separate power constraints at the source and the relay, which is different from the constraint in the previous section where the two constraints is incorporated to be a total power constraint. The last two constraints show that only one-to-one subcarrier matching is permitted, which distinguishes the third way from the first way mentioned.

For evaluation, we transform the above optimization to another one. By introducing the parameter C_i , the optimization problem can be transformed into

$$\begin{aligned}
 & \max_{P_{s,i}, P_{r,j}, \rho_{ij}, C_i} \sum_{i=1}^N C_i \\
 \text{subject to} & \quad \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) \geq C_i \\
 & \quad \sum_{j=1}^N \rho_{ij} \frac{1}{2} \log_2 \left(1 + \frac{P_{r,j} h_{r,j}}{N_0} \right) \geq C_i \\
 & \quad \sum_{i=1}^N P_{s,i} \leq P_s, \sum_{j=1}^N P_{r,j} \leq P_r \\
 & \quad P_{s,i}, P_{r,j} \geq 0, \forall i, j \\
 & \quad \sum_{j=1}^N \rho_{ij} = 1, \rho_{ij} \in \{0, 1\}, \forall i, j
 \end{aligned}$$

That is, the original maximization problem is transformed to a mixed binary integer programming problem. However, it is prohibitive to find the global optimum in terms of computational complexity. In order to determine the optimal solution, an exhaustive search is needed which has been proved to be NP-hard and is fundamentally difficult to solve (Korte & Vygen, 2002). For each subcarrier matching possibility, find the corresponding system capacity, and the largest one is optimal. The corresponding subcarrier matching and power allocation is optimal joint subcarrier matching and power allocation.

In following subsection, by separating subcarrier matching and power allocation, the optimal solution of the above optimization problem is proposed. For the global optimum, the optimal subcarrier matching is proved; then, the optimal power allocation is provided for the optimal subcarrier matching. Additionally, a suboptimal scheme with less complexity is also proposed to better understand the effect of power allocation, and the capacity of suboptimal scheme delivering performance is close to the upper bound of system capacity.

3.2 Optimal subcarrier matching for global optimum

First, the optimal subcarrier matching is provided for system including two subcarriers. Then, the way of optimal subcarrier matching is extended to the system including unlimited number of subcarriers.

3.2.1 Optimal subcarrier matching for the system including two subcarriers

For the mixed binary integer programming problem, the optimal joint subcarrier matching and power allocation can be found by two steps: (1) for every matching possibility (i.e., ρ_{ij} is

given), find the optimal power allocation and the total channel capacity; (2) compare the all channel capacities, the largest one is the ultimate system capacity, whose subcarrier matching and power allocation are jointly optimal. But, this process is prohibitive to find global optimum in terms of complexity. In this subsection, an analytical argument is given to prove that the optimal subcarrier matching is to match subcarrier by the order of the channel power gains.

Here, we assume that the system includes only two subcarriers, i.e., $N = 2$. The channel power gains over the source-relay channel are denoted as $h_{s,1}$ and $h_{s,2}$, and the channel power gains over the relay-destination channel are denoted as $h_{r,1}$ and $h_{r,2}$. Without loss of generality, we assume that $h_{s,1} \geq h_{s,2}$ and $h_{r,1} \geq h_{r,2}$, i.e., the subcarriers are sorted according to the channel power gains. The system power constraints are P_s and P_r at the source and the relay, separately.

In this case, the mixed binary integer programming problem can be reduced to the following optimization problem.

$$\begin{aligned}
 & \max_{P_{s,i}, P_{r,j}, \rho_{ij}, C_i} \sum_{i=1}^2 C_i \\
 & \text{subject to } \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) \geq C_i \\
 & \sum_{j=1}^2 \rho_{ij} \frac{1}{2} \log_2 \left(1 + \frac{P_{r,j} h_{r,j}}{N_0} \right) \geq C_i \\
 & \sum_{i=1}^2 P_{s,i} \leq P_s, \sum_{j=1}^2 P_{r,j} \leq P_r \\
 & P_{s,i}, P_{r,j} \geq 0, \forall i, j \\
 & \sum_{j=1}^2 \rho_{ij} = 1, \rho_{ij} = \{0, 1\}, \forall i, j
 \end{aligned}$$

Here, there are two possibilities to match the subcarriers: (1) the subcarrier 1 over the sourcerelay channel is matched to the subcarrier 1 over the relay-destination channel, and the subcarrier 2 over the source-relay channel is matched to the subcarrier 2 over the relay-destination channel (i.e., $h_{s,1} \sim h_{r,1}$ and $h_{s,2} \sim h_{r,2}$); (2) the subcarrier 1 over the source-relay channel is matched to the subcarrier 2 over the relay-destination channel, and the subcarrier 2 over the source-relay channel is matched to the subcarrier 1 over the relay-destination channel (i.e., $h_{s,1} \sim h_{r,2}$ and $h_{s,2} \sim h_{r,1}$). As there are only two possibilities, the optimal subcarrier matching can be obtained by comparing the capacities of two possibilities. However, the process has to be repeated when the channel power gains are changed. Next, optimal subcarrier matching way will be given without computing the capacities of all subcarrier matching possibilities, after Lemma 2 is proposed and proved.

Lemma 2: For global optimum of the upper optimization problem, the capacity of the better subcarrier is greater than that of the worse subcarrier, where better and worse are according to the channel power gain at the source and the relay.

Proof: We will prove this *Lemma* in the contrapositive form. First, for the global optimum, we assume the power allocations at the source are $P'_{s,1}$ and $P_s - P'_{s,1}$, and assume $R'_{s,1} \leq R'_{s,2}$, i.e., the capacity of better subcarrier is less than that of worse subcarrier, which means

$$\log_2 \left(1 + \frac{h_{s,1} P'_{s,1}}{N_0} \right) \leq \log_2 \left(1 + \frac{h_{s,2} (P_s - P'_{s,1})}{N_0} \right) \quad (27)$$

As the capacity of optimum is the greatest one, the capacity is greater than any other power allocation. When the subcarrier matching is constant, there are no other power allocations to the two subcarriers denoted as $P_{s,1}^*$ and $P_s - P_{s,1}^*$, which make the capacities of two subcarrier satisfied with following relations

$$R_{s,1}^* \geq R'_{s,2} \quad (28)$$

$$R_{s,2}^* \geq R'_{s,1} \quad (29)$$

If the power allocation $P_{s,1}^*$ and $P_s - P_{s,1}^*$ exist, we can rematch the subcarriers to improve system capacity by exchanging the subcarrier 1 and subcarrier 2, i.e., changing the subcarrier matching. According to the new subcarrier matching and power allocation, it is clear that the system capacity can be improved.

Here, we will prove that there exist the power allocations which are satisfied with the equations (28) and (29).

$$\log_2 \left(1 + \frac{h_{s,1} P_{s,1}^*}{N_0} \right) \geq \log_2 \left(1 + \frac{h_{s,2} (P_s - P'_{s,1})}{N_0} \right) \quad (30)$$

$$\log_2 \left(1 + \frac{h_{s,2} (P_s - P_{s,1}^*)}{N_0} \right) \geq \log_2 \left(1 + \frac{h_{s,1} P'_{s,1}}{N_0} \right) \quad (31)$$

By solving the above inequalities, we can get the following inequation

$$\frac{h_{s,2}}{h_{s,1}} (P_s - P'_{s,1}) \leq P_{s,1}^* \leq P_s - \frac{h_{s,1}}{h_{s,2}} P'_{s,1} \quad (32)$$

At the same time, to satisfy the inequality (27), the following relation has to be satisfied

$$P'_{s,1} \leq \frac{h_{s,2} P_s}{h_{s,1} + h_{s,2}} \quad (33)$$

By making use of the above inequality, we can get

$$\begin{aligned} \frac{h_{s,2}}{h_{s,1}} (P_s - P'_{s,1}) - \left(P_s - \frac{h_{s,1}}{h_{s,2}} P'_{s,1} \right) &= \frac{h_{s,2}}{h_{s,1}} P_s - P_s + \frac{(h_{s,1} + h_{s,2})(h_{s,1} - h_{s,2})}{h_{s,1} h_{s,2}} P'_{s,1} \\ &\leq \frac{h_{s,2}}{h_{s,1}} P_s - P_s + \frac{(h_{s,1} + h_{s,2})(h_{s,1} - h_{s,2})}{h_{s,1} h_{s,2}} \frac{h_{s,2} P_s}{h_{s,1} + h_{s,2}} \\ &= \frac{h_{s,2}}{h_{s,1}} P_s - P_s - \frac{h_{s,2}}{h_{s,1}} P_s + P_s \\ &= 0 \end{aligned}$$

Therefore, the following inequality is proved

$$\frac{h_{s,2}}{h_{s,1}}(P_s - P'_{s,1}) \leq P_s - \frac{h_{s,1}}{h_{s,2}} P'_{s,1} \quad (34)$$

This means that we can always find $P_{s,1}^*$ which satisfies the inequality (32). The new power allocation $P_{s,1}^*$ makes the inequalities (28) and (29) satisfied.

Then, we can rematch the subcarriers by exchanging the subcarrier 1 and subcarrier 2 at the source to improve the system capacity. This means that the system capacity of the new subcarrier matching and power allocation is greater than that of the original power allocation.

Therefore, for any power allocations which make the subcarrier capacity of worse subcarrier is greater than that of the better subcarrier, we always can find new power allocation to improve system capacity and make the subcarrier capacity of better subcarrier greater than that of worse subcarrier.

At the relay, for the global optimum, the similar process can be used to prove that the capacity of better subcarrier is greater than that of the worse subcarrier.

Therefore, for the global optimum at the source and the relay, we can conclude that the subcarrier capacity of better subcarrier is greater than that of the worse subcarrier with any channel power gains.

By making use of *Lemma 2*, the following proposition can be proved, which states the optimal subcarrier matching way for the global optimum.

Proposition 4: For the global optimum in the system including only two subcarriers, the optimal subcarrier matching is that the better subcarrier is matched to the better subcarrier and the worse subcarrier is matched to the worse subcarrier, i.e., $h_{s,1} \sim h_{r,1}$ and $h_{s,2} \sim h_{r,2}$.

Proof: Following *Lemma 2*, we know that the capacity of the better subcarrier is greater than the capacity of the worse subcarrier for the global optimum, i.e., $R_{s,1}^* \geq R_{s,2}^*$, $R_{r,1}^* \geq R_{r,2}^*$. There are two ways to match subcarrier: first, the better subcarrier is matched to the better subcarrier, i.e., $h_{s,1} \sim h_{r,1}$ and $h_{s,2} \sim h_{r,2}$; second, the better subcarrier is matched to the worse subcarrier, i.e., $h_{s,1} \sim h_{r,2}$ and $h_{s,2} \sim h_{r,1}$.

We can prove the optimal subcarrier matching is the first way by proving the following inequality

$$\min(R_{s,1}^*, R_{r,1}^*) + \min(R_{s,2}^*, R_{r,2}^*) \geq \min(R_{s,1}^*, R_{r,2}^*) + \min(R_{s,2}^*, R_{r,1}^*) \quad (35)$$

where the left is the system capacity of the first subcarrier matching and the right is that of the second subcarrier matching.

To prove the upper inequality, we can list all possible relations of $R_{s,1}^*$, $R_{r,1}^*$, $R_{s,2}^*$ and $R_{r,2}^*$. Restricted to the relations $R_{s,1}^* \geq R_{s,2}^*$ and $R_{r,1}^* \geq R_{r,2}^*$, there are six possibilities (1) $R_{s,1}^* \geq R_{s,2}^* \geq R_{r,1}^* \geq R_{r,2}^*$; (2) $R_{s,1}^* \geq R_{r,1}^* \geq R_{s,2}^* \geq R_{r,2}^*$; (3) $R_{s,1}^* \geq R_{r,1}^* \geq R_{r,2}^* \geq R_{s,2}^*$; (4) $R_{r,1}^* \geq R_{r,2}^* \geq R_{s,1}^* \geq R_{s,2}^*$; (5) $R_{r,1}^* \geq R_{s,1}^* \geq R_{r,2}^* \geq R_{s,2}^*$; (6) $R_{r,1}^* \geq R_{s,1}^* \geq R_{s,2}^* \geq R_{r,2}^*$. For the every possibility, it is easy to prove the inequality (35) satisfied. Details are omitted for sake of the length.

So far, for the system including two subcarriers, the optimal joint subcarrier matching has been given. Specially, the optimal subcarrier matching is to match the subcarriers by the order of the channel power gains.

3.2.2 Optimal subcarrier matching for the system including unlimited number of subcarriers

This subsection extends the method in the previous subsection to the system including unlimited number of the subcarriers. The number of the subcarriers is finite (e.g., $2 \leq N \leq \infty$), where the subcarrier channel power gains are $h_{s,i}$ and $h_{r,j}$.

As before the channel power gains are assumed $h_{s,i} \geq h_{s,i+1}$ ($1 \leq i \leq N-1$) and $h_{r,j} \geq h_{r,j+1}$ ($1 \leq j \leq N-1$). For the global optimum, the following proposition gives the optimal subcarrier matching.

Proposition 5: For the global optimum in the system including unlimited number of the subcarriers, the optimal subcarrier matching is

$$h_{s,i} \sim h_{r,i} \quad (36)$$

Together with the optimal power allocation for this subcarrier matching, they are optimal joint subcarrier matching and power allocation

Proof: This proposition will be proved in the contrapositive form. For the global optimum, assuming that there is a subcarrier matching method whose matching result including two matched subcarrier pairs $h_{s,i} \sim h_{r,i+n}$ and $h_{s,i+n} \sim h_{r,i}$ ($n > 0$), and the total capacity is greater than that of the matching method in *Proposition 4*.

When the power allocated to other subcarriers and the other subcarrier matching are constant, the total channel capacity of the two subcarrier pairs can be improved based on *Proposition 4*, which implies the channel capacity can be improved by rematching the subcarriers to $h_{s,i} \sim h_{r,i}$ and $h_{s,i+n} \sim h_{r,i+n}$. It is contrary to the assumption. Therefore, there is no subcarrier matching way better than the way in *Proposition 4*. At the same time, as the total capacity of this subcarrier matching and the corresponding optimal power allocation scheme is the largest one, this subcarrier matching together with the corresponding optimal power allocations is the optimal joint subcarrier matching and power allocation.

Therefore, for the system including unlimited number of the subcarriers, the optimal subcarrier matching is to match the subcarrier according to the order of channel power gains, i.e., $h_{s,i} \sim h_{r,i}$. As it is optimal subcarrier matching for the global optimum, together with the optimal power allocation for this subcarrier matching, they are optimal joint subcarrier matching and power allocation.

3.3 Optimal power allocation for optimal subcarrier matching

When the subcarrier matching is given, the parameters ρ_{ij} in optimization problem (9) is constant, e.g., $\rho_{ii} = 1$ and $\rho_{ij} = 0$ ($i \neq j$). Therefore, the optimization problem can be reduced to as follows

$$\begin{aligned} & \max_{P_{s,i}, P_{r,i}, C_i} \sum_{i=1}^N C_i \\ \text{subject to} & \quad \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) \geq C_i \\ & \quad \frac{1}{2} \log_2 \left(1 + \frac{P_{r,i} h_{r,i}}{N_0} \right) \geq C_i \\ & \quad \sum_{i=1}^N P_{s,i} \leq P_s, \sum_{i=1}^N P_{r,i} \leq P_r \\ & \quad P_{s,i}, P_{r,i} \geq 0, \forall i, j \end{aligned}$$

It is easy to prove that the above optimization problem is a convex optimization problem (Boyd & Vanderberghe, 2004). By this way, we have transformed the mixed binary integer programming problem to a convex optimization problem. Therefore, we can solve it to get the optimal power allocation for the optimal subcarrier matching.

Consider the Lagrangian

$$L(\mu_{s,i}, \mu_{r,i}, \gamma_s, \gamma_r) = -\sum_{i=1}^N C_i + \sum_{i=1}^N \mu_{s,i} \left(C_i - \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) \right) + \gamma_s \left(\sum_{i=1}^N P_{s,i} - P_s \right) + \sum_{i=1}^N \mu_{r,i} \left(C_i - \frac{1}{2} \log_2 \left(1 + \frac{P_{r,i} h_{r,i}}{N_0} \right) \right) + \gamma_r \left(\sum_{i=1}^N P_{r,i} - P_r \right)$$

where $\mu_{s,i} \geq 0$, $\mu_{r,i} \geq 0$, $\gamma_s \geq 0$, $\gamma_r \geq 0$ are the Lagrangian parameters.

By making the derivations of $P_{s,i}$ and $P_{r,i}$ equal to zero, we can get the following equations

$$P_{s,i} = \frac{\mu_{s,i}}{2\gamma_s \ln 2} - \frac{N_0}{h_{s,i}} \quad (37)$$

$$P_{r,i} = \frac{\mu_{r,i}}{2\gamma_r \ln 2} - \frac{N_0}{h_{r,i}} \quad (38)$$

By making the derivation of C_i equal to zero, we can get the following equations

$$\mu_{s,i} + \mu_{r,i} = 1 \quad (39)$$

At the same time, for the Lagrangian parameters, we can get the following equations based on KKT conditions (Boyd & Vanderberghe, 2004)

$$\mu_{s,i} \left(C_i - \frac{1}{2} \log_2 \left(1 + \frac{P_{s,i} h_{s,i}}{N_0} \right) \right) = 0 \quad (40)$$

$$\mu_{r,i} \left(C_i - \frac{1}{2} \log_2 \left(1 + \frac{P_{r,i} h_{r,i}}{N_0} \right) \right) = 0 \quad (41)$$

For the summation of subcarrier allocated power at the source and the relay, we make the unequal equation be equal, i.e.,

$$\sum_{i=1}^N P_{s,i} = P_s \quad (42)$$

$$\sum_{i=1}^N P_{r,i} = P_r \quad (43)$$

It is noted that we make the summations of subcarrier power equal to the power constrains at the source and the relay, separately. It is clear that the system capacity will not be reduced by this mechanism.

By making use of the equations (35)-(43), the parameters $\mu_{s,i}$, $\mu_{r,i}$, γ_s and γ_r can be provided. Therefore, the optimal power allocation is achieved. From the expression of power allocation, the power allocation is like based on water-filling. But for different subcarrier, the water surface is different, which is because of the parameters $\mu_{s,i}$ and $\mu_{r,i}$ in power expressions. The power computation is more complex than water-filling algorithm.

In the proof of optimal subcarrier matching, we proved that the optimal subcarrier matching is globally optimal for joint subcarrier matching and power allocation. Therefore, the optimal subcarrier matching is optimal for the optimal power allocation. For optimal joint subcarrier matching and power allocation scheme, it means that the subcarrier matching parameters have to be $\rho_{ii} = 1$ and $\rho_{ij} = 0 (i \neq j)$. Then, the optimal power allocation is obtained according to the globally optimal subcarrier matching parameters. Therefore, the joint subcarrier matching and power allocation scheme is globally optimal. It is different from iterative optimization approach for different parameters where optimization has to be utilized iteratively.

For the system including any number of the subcarriers, the optimal joint subcarrier matching and power allocation scheme has been given by now. Here, the steps are summarized as follows

- Step 1. Sort the subcarriers at the source and the relay in descending order by the permutations π and π' , respectively. The process is according to the channel power gains, i.e., $h_{s,\pi(i)} \geq h_{s,\pi(i+1)}$, $h_{r,\pi'(i)} \geq h_{r,\pi'(i+1)}$.
- Step 2. Match the subcarriers into pairs by the order of the channel power gains (i.e., $h_{s,\pi(i)} \sim h_{r,\pi'(i)}$), which means that the bits transported on the subcarrier $\pi(i)$ over the sourcerelay channel will be retransmitted on the subcarrier $\pi'(i)$ over the relay-destination channel.
- Step 3. Using *Proposition 2*, get the optimal power allocation for the subcarrier matching based on the equations (24) and (25).
- Step 4. According to the optimal joint subcarrier matching and power allocation, get the capacities of all subcarrier at the source and the relay. The capacity of a matched subcarrier pair is

$$C_i = \min \left\{ \frac{1}{2} \log_2 \left(1 + \frac{P_{s,\pi(i)} h_{s,\pi(i)}}{N_0} \right), \frac{1}{2} \log_2 \left(1 + \frac{P_{r,\pi'(i)} h_{r,\pi'(i)}}{N_0} \right) \right\} \quad (44)$$

- Step 5. The total system channel capacity is

$$R_{tot} = \sum_{i=1}^N C_i \quad (45)$$

3.4 The suboptimal scheme

In order to obtain the insight about the effect of power allocation and understand the effect of power allocation, a suboptimal joint subcarrier matching and power allocation is proposed. In optimal scheme, the power allocation is like water-filling but with different water surface at different subcarrier. We infer that the power allocation can be obtained according to water-filling at least at one side. The different power allocation has little effect on the system capacity.

In section 4, the simulations will show that the capacity of optimal scheme is almost equal to the upper bound of system capacity. However, the upper bound is the less one of the capacities of source-relay channel and relay-destination channel. These results motivate us to give the suboptimal scheme. In the suboptimal scheme, the main idea is to make the capacity of the suboptimal scheme as close to the less one as possible of the capacities of source-relay channel and relay-destination channel. Therefore, we hold the power allocation at the less side and make the capacity of the matched subcarrier at the greater side close to the corresponding subcarrier at the less one. At the same time, it is noted that the better subcarrier will need less power than the worse subcarrier to achieve the same capacity improvement. It means that the better subcarrier will have more effect on system capacity by reallocating the power. Therefore, the power reallocation will be made from the best subcarrier to the worst subcarrier at the greater side.

The globally optimal subcarrier matching can be accomplished by simple permutation. Therefore, the same subcarrier matching as the optimal scheme is adopted. The power allocation is different from the optimal scheme. First, to maximize the capacity, we perform water-filling algorithm at the source and the relay separately to get the maximum capacities of source-relay channel and relay-destination channel. In order to close the less one, we keep the power allocation and capacity at the less side, and try to make the greater side equal to the less side. The power reallocation will be made from the best subcarrier to the worst subcarrier at the greater side. Without loss of generality, we assume that the capacity of source-relay channel is less than that of relay-destination channel after applying water-filling algorithm. This means that we keep the power allocation at the source and reallocate power at the relay to make the subcarrier capacity equal to the corresponding subcarrier from the best subcarrier to the worst subcarrier at the relay. Therefore, the less one of them is the capacity of suboptimal scheme. It is noted that the suboptimal scheme still separates the subcarrier matching and power allocation and the subcarrier matching is the same as that of optimal scheme.

The scheme can be described in detail as follows:

- Step 1. Sort the subcarriers at the source and the relay in descending order by permutations π and π' , respectively. The process is according to the channel gains, i.e., $h_{s,\pi(i)} \geq h_{s,\pi(i+1)}$, $h_{r,\pi'(j)} \geq h_{r,\pi'(j+1)}$. Then, match the subcarriers into pairs at the same order of both nodes (i.e., $\pi(k) \sim \pi'(k)$), which means that the bits transported on the subcarrier $\pi(k)$ at the source will be retransmitted on the subcarrier $\pi'(k)$ at the relay.
- Step 2. Perform the water-filling algorithm to get the respective channel capacity at the source and the relay. Without loss of generality, we assume the channel capacity over source-relay channel is less than the total channel capacity over relay-destination channel.
- Step 3. From $k = 1$ to N , reallocate the power to subcarrier $\pi'(k)$ so that $R_{r,\pi'(k)} = R_{s,\pi(k)}$ until $\sum_{i=1}^k P_{r,\pi'(i)} \geq P_r$ or $k = N$. The power allocated to the k th subcarrier is $P_r - \sum_{i=1}^k P_{r,\pi'(i)}$ if $k < N$ and $\sum_{i=1}^k P_{r,\pi'(i)} \geq P_r$, and the power allocated to the other subcarriers is zero.

The power allocation of the suboptimal scheme includes performing water-filling algorithm twice and some line operations, which is easier than that of optimal joint subcarrier matching and power allocation. Next, the simulations will prove that the capacity of

suboptimal is close to that of optimal scheme. The main reasons include two: (1) The subcarrier matching of the suboptimal scheme is globally optimal as that of the optimal scheme. (2) The method of power allocation is to make the capacity as close to the upper bound as possible. The subcarrier with more effect on the capacity is considered firstly through power allocation.

4. Simulation

In this section, the capacities of the optimal and suboptimal schemes are compared with that of several other schemes and the upper bound of system capacity with separate power constraints by computer simulations. These schemes include:

- i. No subcarrier matching and no water-filling with separate power constraints: the bits transmitted on the subcarrier i at the source will be retransmitted on the subcarrier i at the relay; the power is allocated equally among the all subcarriers at the source and the relay, separately. It is denoted as *no matching & no water-filling* in the figures.
- ii. Water-filling and no subcarrier matching with separate power constraints: the bits transmitted on the subcarrier i at the source will be retransmitted on the subcarrier i at the relay; the power allocation is according to water-filling at the source and the relay, separately. It is denoted as *water-filling & no matching* in the figures.
- iii. Subcarrier matching and no water-filling with separate power constraints: the bits transmitted on the subcarrier $\pi(i)$ at the source will be retransmitted on the subcarrier $\pi'(i)$ at the relay; the power is allocated equally among the all subcarriers at the source and the relay, separately. It is denoted as *matching & no water-filling* in the figures.
- iv. Subcarrier matching and water-filling with separate power constraints: the bits transmitted on subcarrier $\pi(i)$ at the source will be retransmitted on the subcarrier $\pi'(i)$ at the relay; the power is allocated according to water-filling algorithm at the source and the relay, separately. It is denoted as *matching & water-filling* in the figures.
- v. Optimal joint subcarrier matching and power allocation with total power constraint. Here, the power constraint is system-wide. It is denoted as *optimal & total* in the figures. Here, the subcarrier matching is the same as that of optimal and suboptimal schemes, which can be complemented according to the *Step 1 - Step 2* in the optimal scheme. The water-filling means that the water-filling algorithm is performed at the source and the relay only once.

According to the complexity, the suboptimal scheme has less complexity than the optimal scheme, where the difference comes from different power allocation. For the optimal scheme, the optimal power allocation is like based on water-filling, which can be obtained by multiwaterlevel water-filling solution with complexity $\mathcal{O}(2n)$ according to the reference (Palomar & Fonollosa, 2005). The power allocation of suboptimal scheme can be obtained by water-filling and some linear operation with complexity $\mathcal{O}(n)$ according to the reference (Palomar & Fonollosa, 2005). Therefore, the suboptimal has less complexity than optimal scheme. The other schemes without power allocation or subcarrier matching have less complexity compared with the optimal and suboptimal schemes.

In the computer simulations, it is assumed that each subcarrier undergoes identical Rayleigh fading independently and the average channel power gains, $E(h_{s,i})$ and $E(h_{r,i})$ for all i and j , are assumed to be one. Though the Rayleigh fading is assumed, it is noted that the proof of optimal subcarrier matching utilizes only the order of the subcarrier channel power gains.

The concrete fading distribution has nothing to do with the optimal subcarrier matching. The optimal power allocation for the optimal subcarrier is not utilizing the Rayleigh fading assumption. Therefore, the proposed scheme is effective for other fading distribution, and the same subcarrier matching and power allocation scheme can be adopted. The total bandwidth is $B = 1\text{MHz}$. The SNR_s is defined as $P_s/(N_0B)$ and SNR_r is defined as $P_r/(N_0B)$. To obtain the average data rate, we have simulated 10,000 independent trials.

Fig. 4 shows the capacity versus $SNR_s = SNR_r$. In Fig.4, for the system with separate power constraints, it is noted that the capacity of optimal scheme is approximately equal to upper bound of capacity, which proves that the one-to-one subcarrier matching is approximately optimal. Furthermore, the one-to-one subcarrier matching simplifies the system architecture. The capacity of suboptimal scheme is also close to that of optimal scheme. This can be explained by the approximate equality of capacity of suboptimal scheme to the upper bound of system capacity. Meanwhile, it is also noted that the capacity of suboptimal scheme is greater than that of *subcarrier matching & water-filling*. Though the power allocations at the less side of the two schemes are in same way, the power reallocation at the greater side can improve the system capacity for the suboptimal scheme. The reason is that the capacity of the matched subcarrier over the greater side may be less than that of the corresponding subcarrier over the less side, and limit the capacity of the matched subcarrier pair. However, it is avoided in the suboptimal scheme by power reallocation at the greater side. Another result is that the capacities of optimal and suboptimal schemes are higher than that of other schemes. If there is no subcarrier matching, power allocation by water-filling algorithm decreases the system capacity, which can be obtained by comparing the capacity of

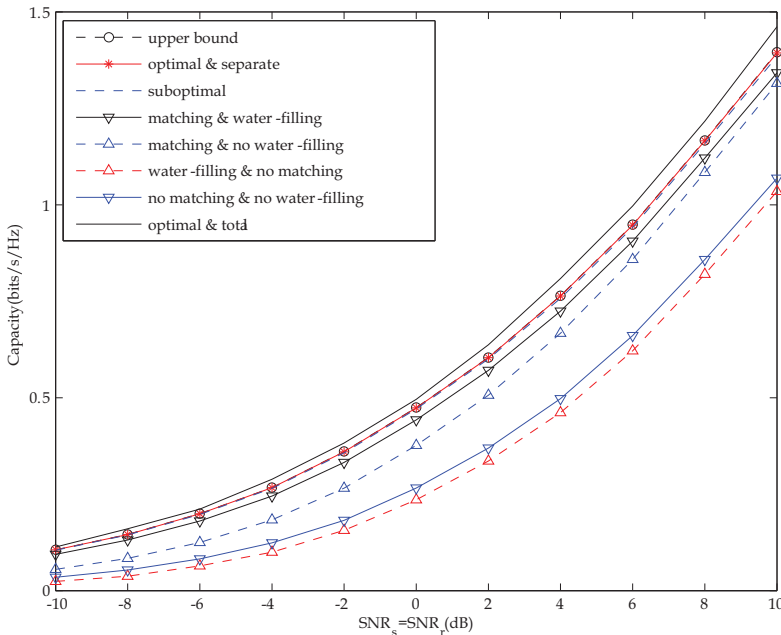


Fig. 4. Channel capacity against $SNR_s = SNR_r (N = 16)$

scheme (i) to that of scheme (ii). The reason is that the water-filling can amplify the capacity imbalance between that of the subcarriers of matched subcarrier pair. For example, when a better subcarrier is matched to a worse subcarrier, the capacity of the matched subcarrier pair is greater than zero with equal power allocation. But the capacity may be zero with water-filling because the worse subcarrier may have no allocated power according to water-filling. The subcarrier matching can improve capacity by comparing the capacity of scheme (i) to that of scheme (iii). However, when only one method is permitted to be used to improve capacity, the subcarrier matching is preferred, which can be obtained by comparing the capacity of scheme (ii) to that of scheme (iii). When $SNR_s = SNR_r$, the capacity of optimal scheme with total power constraint is greater than that of optimal scheme with separate power constraints. Though $SNR_s = SNR_r$ in the system with separate power constraints, the different channel power gains of subcarriers can still lead to different capacities of the source-relay channel and the relay-destination channel. The less one will still limit the system capacity. When the system has the total power constraints, the power allocation can be always found to make the capacities of source-relay channel and relay-destination channel equal to each other. It can avoid the capacity imbalance between that of source-relay channel and relay-destination channel, and improve the system capacity.

The relation between the system capacity and SNR at the source is shown in Fig.5, where the SNR at the relay is constant. The SNR difference may be caused by the different distance at source-relay and relay-destination or different power constraint at the source and the relay. Here, for the system with separate power constraints, the capacity of optimal scheme is still almost equal to the upper bound of capacity and the capacity of suboptimal scheme is still close to that of optimal scheme. The greater is the SNR difference between the source and the relay, the smaller is the difference between the optimal scheme and suboptimal scheme. This proves that the suboptimal scheme is effective. The capacities of optimal and suboptimal schemes are still higher than that of other schemes. When the SNR difference is great between the source and the relay, the capacity of scheme (i) is close to the scheme (ii). It is because of the power allocation has less effect on the difference of subcarrier capacity. But, the subcarrier matching always can improve system capacity with any SNR difference between the source and the relay. It is also noted the capacity of optimal scheme with total power constraint is always improving with the SNR at the source. The reason is that total power be increased as the power at the source.

In order to evaluate the effect of the different power constraint at the source and the relay, the relations between the system capacity and SNR at the relay is also shown in Fig.6. Almost same results as those shown in the Fig.5 can be obtained by exchanging the role of SNR at the source and that at the relay. For the system with separate power constraints, the capacity of optimal scheme is still almost equal to the upper bound of system capacity and the capacity of suboptimal scheme is still close to that of optimal scheme. The greater is the SNR difference between the source and the relay, the smaller is the difference between the optimal scheme and suboptimal scheme. This prove that the suboptimal scheme is effective. The capacities of optimal and suboptimal schemes are still higher than that of other schemes. When the SNR difference is great between the source and the relay, the capacity of scheme (i) is close to the scheme (ii). It is because of the power allocation has little effect on the difference of subcarrier capacity with great SNR difference. But, the subcarrier matching can always increase system capacity with any SNR difference between the source and the

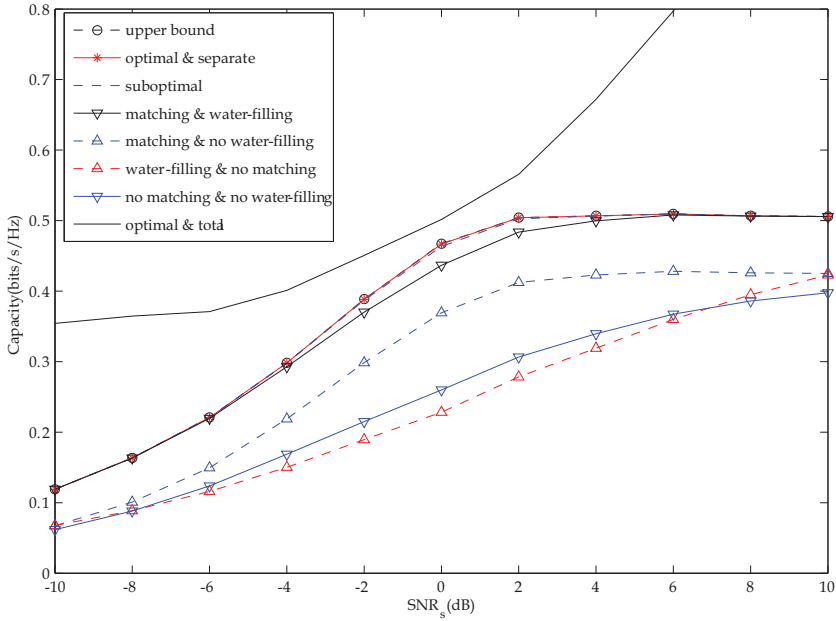


Fig. 5. Channel capacity against SNR_s ($SNR_r = 0dB, N = 16$)

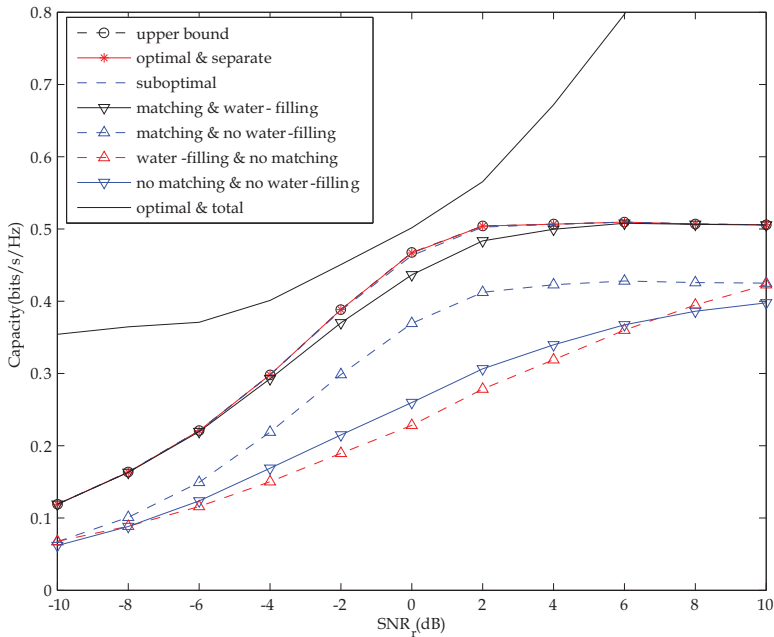


Fig. 6. Channel capacity against SNR_r ($SNR_s = 0dB, N = 16$)

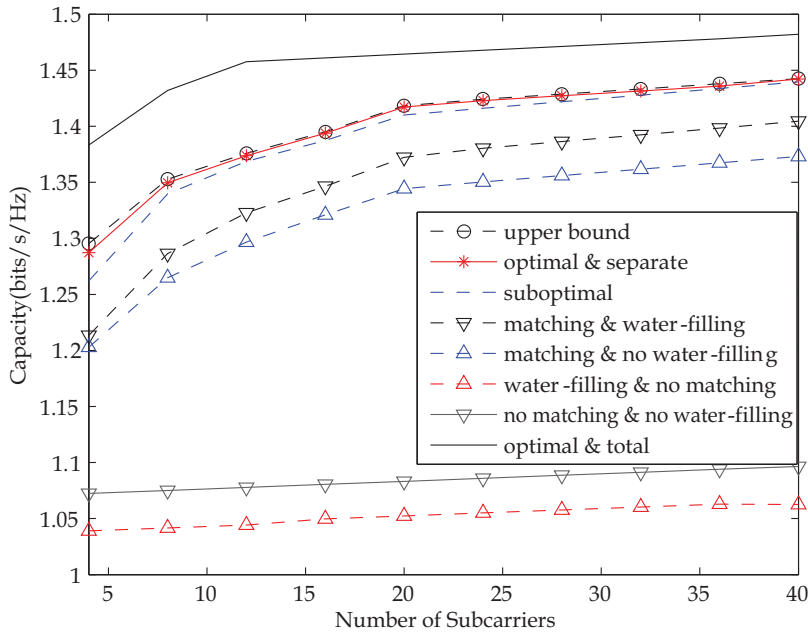


Fig. 7. Channel capacity against the number of subcarriers ($SNR_s = SNR_r = 10dB$)

relay. It is also noted the capacity of optimal scheme with total power constraint is always improved with increasing of the SNR at the source. The reason is that total power will be improved with the power at the relay. The similarity between the Fig. 5 and Fig. 6 proves that the power constraints at the source and the relay have similar effect on the system capacity. It is because that the system capacity will be limited by any less capacity between that of the source-relay channel and the relay-destination channel. When the any node has the less power, the corresponding capacity over the channel will be less than the other and limit the system capacity.

The relation between the system capacities and the number of subcarriers is shown in Fig.7, where the $SNR_s = SNR_r = 10dB$. According to the comparisons among the schemes, similar conclusions can be obtained. With the increasing of number of subcarriers, the system capacity is increasing slowly, which is because of the constant total bandwidth and SNR. For the any number of subcarriers, the capacity of *optimal & total* is greater than that of *optimal & separate*. For the total power constraint, the power can be allocated between the source and the relay, which can avoid the capacity imbalance between that of source-relay channel and relay-destination channel.

In conclusion, the capacity of optimal scheme is approximately equal to the upper bound of system capacity at any circumstance. Therefore, we can always simply the system architecture by only one-to-one subcarrier matching and careful power allocation.

5. Conclusion

The resource allocation problem has been discussed, i.e., joint subcarrier matching and power allocation, to maximize the system capacity for OFDM two-hop relay system. Though the

optimal joint subcarrier matching and power allocation problem is a binary mixed integer programming problem and prohibitive to find global optimum, the optimal joint subcarrier matching and power allocation schemes are provided by separating the subcarrier matching and power allocation. For the global optimum, the optimal subcarrier matching is to match subcarrier according to the channel power gains of subcarriers. The optimal power allocation for the optimal subcarrier matching can be obtained by solving a convex optimization problem. For the system with separate power constraints, the capacity of optimal scheme is almost close to the upper bound of system capacity, which prove that one-to-one subcarrier matching is approximately optimal. The simulations shows that the optimal schemes increase the system capacity by comparing them with several other schemes, where there is no subcarrier matching or power allocation.

6. References

- A. Sendonaris, E. Erkip and B. Aazhang (2003), "User cooperation diversity - Part I and II," *IEEE Transactions Communication*, vol. 51, no. 11, pp. 1927-1948.
- G. Kramer, M. Gastpar, P. Gupta(2006), "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037-3063.
- J. N. Laneman, D. N. C. Tse, and G. W. Wornell (2004), "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062- 3080.
- N. Shastry and R. S. Adve (2005), "A theoretical analysis of cooperative diversity in wireless sensor networks," *IEEE Global Telecommunications Conference, GlobeCom'05*, vol. 6, pp. 3269-3273.
- S. Serbetli, A. Yener (2006), "Power Allocation and Hybrid Relaying Strategies for F/TDMA Ad Hoc Networks," *IEEE International Conference on Communications, ICC'06*, vol. 4, pp. 1562-1567.
- G. Li, J. Liu (2004), "On the capacity of broadband relay networks," *Proc. Asilomar Conf. on Signals, Systems and Computers*, pp. 1318-1322.
- H. Zhu, T. Himsoon, W. P. Siriwongpairat, K. J. R. Liu (2005), "Energy-efficient cooperative transmission over multiuser OFDM networks: who helps whom and how to cooperate," *IEEE Wireless Communications and Networking Conference, WCNC'05*, vol. 2, pp. 1030-1035.
- L. Dai, B. Gui and L. J. C., Jr. (2007), "Selective Relaying in OFDM multihop cooperative networks," *IEEE Wireless Communications and Networking Conference, WCNC '07*, pp. 963-968.
- M. Kaneko and P. Popovski. (2007), "Radio resource allocation algorithm for relay-aided cellular OFDMA system," *IEEE International Conference on Communications, ICC'07*, pp. 4831-4836.
- M. Herdin (2006), "A chunk based OFDM amplify-and-forward relaying scheme for 4G mobile radio systems," *IEEE International Conference on Communications, ICC'06*, pp. 4507 -4512.
- T. Cover, J. Thomas (1991), *Elements of Information Theory*, John Wiley & Sons, Inc., New York.

- B. Korte, J. Vygen (2002), *Combinatorial Optimization: Theory and Algorithms*, 3rd ed. New York: Springer-Verlag.
- I. Hammerstrom and A. Wittneben (2006), "On the Optimal Power Allocation for Nonregenerative OFDM Relay Links," *IEEE International Conference on Communications, ICC'06*, pp. 4463-4468.
- B. Gui and L. J. Cimini, Jr. (2008), "Bit Loading Algorithms for Cooperative OFDM Systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, Article ID 476797, 9 pages.
- Y. Ma, N. Yi, and R. Tafazolli (2008), "Bit and Power Loading for OFDM-Based Three-Node Relaying Communications," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3236-3247.
- S. -J. Kim, X. Wang, and M. Madhian (2008), "Optimal Resource Allocation in Multi-hop OFDMA Wireless Networks with Cooperative Relay," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1833-1838.
- M. Pischella, and J. -C. Belfiore (2008), "Power Control in Distributed Cooperative OFDMA Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1900-1906.
- H. A. Suraweera, and J. Armstrong (2007), "Performance of OFDM-Based Dual-Hop Amplify-and- Forward Relaying," *IEEE Communications Letter*, vol. 11, no. 9, pp. 726-728.
- C. Athaudage, M. Saito, and J. Evans (2008), "Performance Analysis of Dual-Hop OFDM Relay Systems with Subcarrier Mapping, " *IEEE International Conference on Communications, ICC'08*.
- W. Wang, S. Yan, and S. Yang (2008). "Optimally Joint Subcarrier Matching and Power Allocation in OFDM Multihop System," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 241378, 8 pages.
- W. Wang, and R. Wu (2009). "Capacity Maximization for OFDM Two-Hop Relay System With Separate Constraints," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 9, pp. 4943-4954.
- A. Pandharipande, and Chin K. Ho (2007). "Spectrum pool reassignment for a cognitive OFDM-based relay system," *2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrownCom'2007*, pp. 90-94.
- A. Pandharipande, and Chin K. Ho (2008). "Spectrum pool reassignment for wireless multihop relay systems," *3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CrownCom'2008*, pp. 1-5.
- S. Boyd and L. Vanderberghe (2004), *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press.
- D. P. Palomar and J. R. Fonollosa (2005), "Practical algorithms for a family of waterfilling solutions," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 686-695.

MC-CDMA Systems: a General Framework for Performance Evaluation with Linear Equalization

Barbara M. Masini¹, Flavio Zabini¹ and Andrea Conti^{1,2}

¹IEIIT/CNR, WiLab and University of Bologna

²ENDIF, University of Ferrara
Italy

1. Introduction

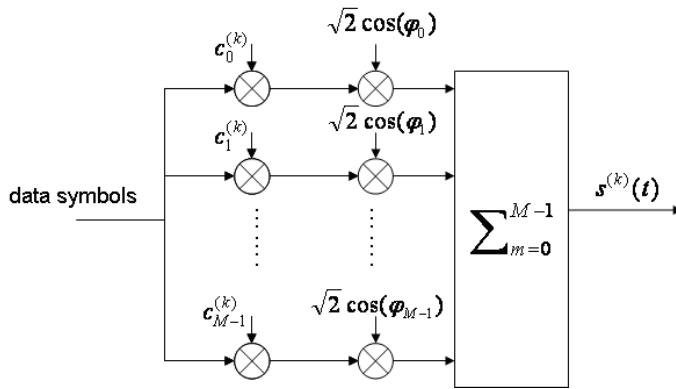
The adaptation of wireless technologies to the users rapidly changing demands is one of the main drivers of the wireless access systems development. New high-performance physical layer and multiple access technologies are needed to provide high speed data rates with flexible bandwidth allocation, hence high spectral efficiency as well as high adaptability.

Multi carrier-code division multiple access (MC-CDMA) technique is candidate to fulfil these requirements, answering to the rising demand of radio access technologies for providing mobile as well as nomadic applications for voice, video, and data. MC-CDMA systems, in fact, harness the combination of orthogonal frequency division multiplexing (OFDM) and code division multiple access (CDMA), taking advantage of both the techniques: OFDM multi-carrier transmission counteracts frequency selective fading channels and reduces signal processing complexity by enabling equalization in the frequency domain, whereas CDMA spread spectrum technique allows the multiple access using an assigned spreading code for each user, thus minimizing the multiple access interference (MAI) (K. Fazel, 2003; Hanzo & Keller, 2006). The advantages of multi-carrier modulation on one hand and the flexibility offered by the spread spectrum technique on the other hand, let MC-CDMA be a candidate technique for next generation mobile wireless systems where spectral efficiency and flexibility are considered as the most important criteria for the choice of the air interface.

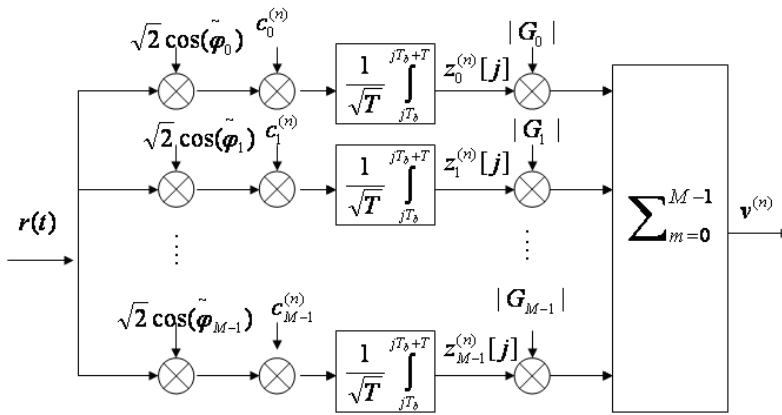
Two different spreading techniques exist, referred to as MC-CDMA (or OFDM-CDMA) with spreading performed in the frequency domain, and MC-DS-CDMA, where DS stands for direct sequence and the spreading is intended in the time domain.

We consider MC-CDMA systems where the data of different users are spread in the frequency-domain using orthogonal code sequences, as shown in Fig. 1: each data symbol is copied on the overall sub-carriers or on a subset of them and multiplied by a chip of the spreading code assigned to the specific user.

The spreading in the frequency domain allows simple methods of signal detection; in fact, since the fading on each sub-carriers can be considered flat, simple equalization with one complex-valued multiplication per sub-carrier can be realized. Furthermore, since the spreading code length does not have to be necessarily chosen equal to the number of sub-carriers, MC-CDMA structure allows flexibility in the system design (K. Fazel, 2003).



(a) Transmitter block scheme ($\varphi_m = 2\pi f_m t + \phi_m, m = 0 \dots M-1$).



(b) Receiver block scheme ($\tilde{\varphi}_m = 2\pi f_m t + \vartheta_m, m = 0 \dots M-1$).

Fig. 1. Transmitter and receiver block schemes.

2. Equalization techniques

The main impairment of this multiplexing technique is given by the MAI, which occurs in the presence of multipath propagation due to loss of orthogonality among the received spreading codes. In conventional MC-CDMA systems, the mitigation of MAI is accomplished at the receiver by employing single-user or multiuser detection schemes. In fact, the exploitation of suitable equalization techniques at the transmitter or at the receiver, can efficiently combine signals on different sub-carriers, toward system performance improvement.

We focus on the downlink of MC-CDMA systems and, after an overall consideration on general combining techniques, we consider linear equalization, representing the simplest and cheapest techniques to be implemented (this can be relevant in the downlink where the receiver is in the user terminal). The application of orthogonal codes, such as Walsh-

Hadamard (W-H) codes for a synchronous system (e.g., the downlink of a cellular system) guarantees the absence of MAI in an ideal channel and a minimum MAI in real channels.¹

2.1 Linear equalization

Within linear combining techniques, various schemes based on the channel state information (CSI) are known in the literature, where signals coming from different sub-carriers are weighted by suitable coefficients G_m (m being the sub-carrier index).

The equal gain combining (EGC) consists in equal weighting of each sub-carrier contribution and compensating only the phases as in (1)

$$G_m = \frac{H_m^*}{|H_m|} \quad (1)$$

where G_m indicates the m^{th} complex channel gain and H_m is the m^{th} channel coefficient (operation * stands for complex conjugate).

If the number of active users is negligible with respect to the number of sub-carriers, that is the system is noise-limited, the best choice is represented by a combination in which the sub-carrier with higher signal-to-noise ratio (SNR) has the higher weight, as in the maximal ratio combining (MRC)

$$G_m = H_m^* \quad (2)$$

The MRC destroys the orthogonality between the codes. For this reason, when the number of active user is high (the system is interference-limited) a good choice is given by restoring at the receiver the orthogonality between the sequences. This means to cancel the effects of the channel on the sequences as in the orthogonality restoring combining (ORC), also known as zero forcing, where

$$G_m = \frac{1}{H_m} \quad (3)$$

This implies a total cancellation of the multiuser interference, but, on the other hand, this method enhances the noise, because the sub-carriers with low SNR have higher weights. Consequently, a correction on G_m is introduced with threshold orthogonality restoring combining (TORC)

$$G_m = u(|H_m| - \rho_{\text{TH}}) \frac{1}{H_m} \quad (4)$$

where $u(\cdot)$ is the unitary-step function and the threshold ρ_{TH} is introduced to cancel the contributions of sub-carriers highly corrupted by the noise.

However, exception made for the two extreme cases of one active user (giving MRC) and negligible noise (giving ORC) the presented methods do not represent the optimum solution for real cases of interest.

¹ In the uplink a set of spreading codes, such as Gold codes, with good auto- and cross-correlation properties, should be employed. However in this case a multi-user detection scheme in the receiver is essential because the asynchronous arrival times destroy orthogonality among the sub-carriers.

The optimum choice for linear equalization is the minimum mean square error (MMSE) technique, whose coefficient can be written as

$$G_m = \frac{H_m^*}{|H_m|^2 + \frac{1}{N\bar{\gamma}}} \quad (5)$$

where N_u is the number of active users and $\bar{\gamma}$ is the mean SNR averaged over small-scale fading. Hence, in addition to the CSI, MMSE requires the knowledge of the signal power, the noise power, and the number of active users, thus representing a more complex linear technique to be implemented, especially in the downlink, where the combination is typically performed at the mobile unit.

To overcome the additional complexity due to estimation of these quantities, a low-complex suboptimum MMSE equalization can be realized (K. Fazel, 2003). With suboptimum MMSE, the equalization coefficients are designed such that they perform optimally only in the most critical cases for which successful transmission should be guaranteed

$$G_m = \frac{H_m^*}{|H_m|^2 + \lambda} \quad (6)$$

where λ is the threshold at which the optimal MMSE equalization guarantees the maximum acceptable bit error probability (BEP) and requires only information about H_m . However, the value of λ has to be determined during the system design and varies with the scenario.

A new linear combining technique has been recently proposed, named partial equalization (PE), whose coefficient G_m is given by (Conti et al., 2007)

$$G_m = \frac{H_m^*}{|H_m|^{1+\beta}} \quad (7)$$

where β is the PE parameter having values in the range of $[-1,1]$. It may be observed that, being parametric with β , (7) reduces to EGC, MRC and ORC for $\beta = 0, -1$, and 1 , respectively. Hence, (7) includes in itself all the most commonly adopted linear combining techniques.

Note also that, while MRC, and ORC are optimum in the extreme cases of noise-limited and interference-limited systems, respectively, for each intermediate situation an optimum value of the PE parameter β can be found to optimize the performance. Moreover, the PE scheme has the same complexity of EGC, MRC, and ORC, but it is more robust to channel impairments and to MAI-variations (Conti et al., 2007).

2.2 Non-linear equalization

Linear equalization techniques compensate the distortion due to flat fading, by simply performing one complex-valued multiplication per sub-carrier. If the spreading code structure of the interfering signals is known, the MAI could not be considered in advance as noise-like, yielding to suboptimal performance.

Non-linear multiuser equalizers, such as interference cancellation (IC) and maximum likelihood (ML) detection, exploit the knowledge of the interfering users' spreading codes in the detection process, thus improving the performance at the expense of higher receiver complexity (Hanzo et al., 2003).

IC is based on the detection of the interfering users' information and its subtraction from the received signal before the determination of the desired user's information. Two kinds of IC techniques exist: parallel and successive cancellation. Combinations of parallel and successive IC are also possible. IC works in several iterations: each detection stage exploits the decisions of the previous stage to reconstruct the interfering contribution in the received signal. It can be typically applied in cellular radio systems to reduce intra-cell and inter-cell interference. Note that IC requires a feedback component in the receiver and the knowledge of which users are active.

The ML detection attains better performance since it is based on optimum maximum likelihood detection algorithms which optimally estimate the transmitted data. Many optimum ML algorithms have been presented in literature and we remind the reader to (Hanzo et al., 2003; K. Fazel, 2003) for further investigation which are out of the scope of the present chapter. However, since the complexity of ML detection grows exponentially with the number of users and the number of bits per modulation symbol, its use can be limited in practice to applications with few users and low order modulation. Furthermore, also in this case as for IC, the knowledge about which users are active is necessary to compute the possible transmitted sequences and apply ML criteria.

2.3 Objectives of the chapter

We propose a general and parametric analytical framework for the performance evaluation of the downlink of MC-CDMA systems with PE.²In particular,

- we evaluate the performance in terms of bit error probability (BEP);
- we derive the optimum PE parameter β for all possible number of sub-carriers, active users, and for all possible values of the SNR;
- we show that PE technique with optimal β improves the system performance still maintaining the same complexity of MRC, EGC and ORC and is close to MMSE;
- we consider a combined equalization (CE) scheme jointly adopting PE at both the transmitter and the receiver and we investigate when CE introduces some benefits with respect to classical single side equalization.

3. System model

We focus on PE technique, that being parametric includes previously cited linear techniques and allows the derivation of a general framework to assess the performance evaluation and sensitivity to system parameters.

3.1 Transmitter

Referring to binary phase shift keying (BPSK) modulation and to the transmitter block scheme depicted in Fig. 1(a), the transmitted signal referred to the k^{th} user, can be written as

$$s^{(k)}(t) = \sqrt{\frac{2E_b}{M}} \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} c_m^{(k)} a^{(k)}[i] g(t - iT_b) \cos(\varphi_m) \quad (8)$$

² Portions reprinted with permission from A. Conti, B. M. Masini, F. Zabini, and O. Andrisano, "On the down-link Performance of Multi-Carrier CDMA Systems with Partial Equalization", IEEE Transactions on Wireless Communications, Volume 6, Issue 1, Jan. 2007, Page(s):230 - 239. ©2007 IEEE, and from B. M. Masini, A. Conti, "Combined Partial Equalization for MC-CDMA Wireless Systems", IEEE Communications Letters, Volume 13, Issue 12, December 2009 Page(s):884 - 886. ©2009 IEEE.

where E_b is the energy per bit, i denotes the data index, m is the sub-carrier index, c_m is the m^{th} chip (taking value ± 1)³, $a_i^{(k)}$ is the data-symbol transmitted during the i^{th} time-symbol, $g(t)$ is a rectangular pulse waveform, with duration $[0, T]$ and unitary energy, T_b is the bit-time, $\varphi_m = 2\pi f_m t + \phi_m$ where $f_m = f_0 + m \cdot \Delta f$ is the sub-carrier-frequency (with $\Delta f \cdot T$ and $f_0 T$ integers to have orthogonal frequencies) and ϕ_m is the random phase uniformly distributed within $[-\pi, \pi]$. In particular, $T_b = T + T_g$ is the total OFDM symbol duration, increased with respect to T of a time-guard T_g (inserted between consecutive multi-carrier symbols to eliminate the residual inter symbol interference, ISI, due to the channel delay spread). Note that we assume rectangular pulses for analytical purposes. However, this does not lead the generality of the work. In fact, a MC-CDMA system is realized, in practice, through inverse fast Fourier transform (IFFT) and FFT at the transmitter and receiver, respectively. After the sampling process, the signal results completely equivalent to a MC-CDMA signal with rectangular pulses in the continuous time-domain. Considering that, exploiting the orthogonality of the code, all the different users use the same carriers, the total transmitted signal results in

$$s(t) = \sum_{k=0}^{N_u-1} s^{(k)}(t) = \sqrt{\frac{2E_b}{M}} \sum_{k=0}^{N_u-1} \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} c_m^{(k)} a^{(k)}[i] g(t - iT_b) \cos(\varphi_m) \quad (9)$$

where N_u is the number of active users and, because of the use of orthogonal codes, $N_u \leq M$.

3.2 Channel model

Since we are considering the downlink, focusing on the n^{th} receiver, the information associated to different users experiences the same fading. Due to the CDMA structure of the system, each user receives the information of all the users and select only its own data through the spreading sequence. We assume the impulse response of the channel $h(t)$ as time-invariant during many symbol intervals.

We employ a frequency-domain channel model in which the transfer function, $H(f)$, is given by

$$H(f) = H(f_m) = \alpha_m e^{j\varphi_m} \text{ for } |f - f_m| < \frac{W_s}{2}, \forall m \quad (10)$$

where α_m and φ_m are the m^{th} amplitude and phase coefficients, respectively, and W_s is the transmission bandwidth of each sub-carrier. The assumption in (10) means that the pulse shaping still remains rectangular even if the non-distortion conditions are not perfectly verified. Hence, the response $g'(t)$ to $g(t)$ is a rectangular pulse with unitary energy and duration $T' \triangleq T + T_d$, being $T_d \leq T_g$ the time delay. Note that this assumption is helpful in the analytical process and does not impact in the generality of the work.

We assume that each $H(f_m)$ is independent identically distributed (i.i.d.) complex zero-mean Gaussian random variable (r.v.) with variance, σ_{H}^2 , related to the path-loss L_p as $1/L_p = \mathbb{E}\{\alpha^2\} = \sigma_{H}^2$.

³ We assume orthogonal sequences $\overline{c^{(k)}}$ for different users, such that:

$$\langle \overline{c^{(k)}}, \overline{c^{(k')}} \rangle = \sum_{m=0}^{M-1} c_m^{(k)} c_m^{(k')} = \begin{cases} M & k = k' \\ 0 & k \neq k' \end{cases}$$

3.3 Receiver

The received signal can be written as

$$r(t) = \sqrt{\frac{2E_b}{M}} \sum_{k=0}^{N_u-1} \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} \alpha_m c_m^{(k)} a^{(k)}[i] g'(t - iT_b) \cos(\tilde{\varphi}_m) + n(t) \quad (11)$$

where $n(t)$ is the additive white Gaussian noise with two-side power spectral density (PSD) $N_0/2$, $\tilde{\varphi}_m = 2\pi f_m t + \vartheta_m$, and $\vartheta_m \triangleq \phi_m + \psi_m$. Note that, since ϑ_m can be considered uniformly distributed in $[-\pi, \pi]$, we can consider $\angle H(f_m) \sim \vartheta_m$ in the following.

The receiver structure is depicted in Fig. 1(b). Focusing, without loss of generality, to the l^{th} sub-carrier of user n , the receiver performs the correlation at the j^{th} instant (perfect synchronization and phase tracking are assumed) of the received signal with the signal $c_l^{(n)} \sqrt{2} \cos(\tilde{\varphi}_l)$, as

$$z_l^{(n)}[j] = \frac{1}{\sqrt{T}} \int_{jT_b}^{jT_b+T} r(t) c_l^{(n)} \sqrt{2} \cos(\tilde{\varphi}_l) dt. \quad (12)$$

Substituting (11) in (12), the term $z_l^{(n)}[j]$ results in (13)

$$\begin{aligned} z_l^{(n)}[j] &= 2\sqrt{\frac{E_b}{MT}} \sum_{i=-\infty}^{+\infty} \int_{jT_b}^{jT_b+T} \sum_{k=0}^{N_u-1} \sum_{m=0}^{M-1} \alpha_m c_m^{(k)} c_l^{(n)} a^{(k)}[i] g'(t - iT_b) \\ &\quad \times \cos(\tilde{\varphi}_m) \cos(\tilde{\varphi}_l) dt + \overbrace{\int_{jT_b}^{jT_b+T} \sqrt{2} \frac{c_l^{(n)}}{\sqrt{T}} n(t) \cos(\tilde{\varphi}_l) dt}^{n_l[j]} \\ &= \sqrt{\frac{E_b \delta_d}{M}} \alpha_l a^{(n)}[j] + \sqrt{\frac{E_b \delta_d}{M}} c_l^{(n)} \alpha_l \sum_{k=0, k \neq n}^{N_u-1} c_l^{(k)} a^{(k)}[j] + n_l[j] \end{aligned} \quad (13)$$

where $\delta_d \triangleq 1/(1 + T_d/T)$ represents the loss of energy caused by the time-spreading of the impulse.

4. Decision variable

The decision variable, $v^{(n)}[j]$, is obtained by linearly combining the weighted signals from each sub-carrier as follows⁴

$$v^{(n)} = \sum_{l=0}^{M-1} |G_l| z_l^{(n)} \quad (14)$$

where $|G_l|$ is a suitable amplitude of the l^{th} equalization coefficient. By considering PE, the weight for the l^{th} sub-carrier is given by

⁴ For the sake of conciseness in our notation, since ISI is avoided, we will neglect the time-index j in the following.

$$G_l = \frac{H^*(f_l)}{|H^*(f_l)|^{1+\beta}}, \quad -1 \leq \beta \leq 1. \quad (15)$$

Therefore, from (13) and (14) we can write

$$v^{(n)} = \sqrt{\frac{E_b \delta_d}{M}} \sum_{l=0}^{M-1} \alpha_l^{1-\beta} a^{(n)} + \sum_{l=0}^{M-1} \alpha_l^{-\beta} n_l + \sqrt{\frac{E_b \delta_d}{M}} \sum_{l=0}^{M-1} \sum_{k=0, k \neq n}^{N_l-1} \alpha_l^{1-\beta} c_l^{(n)} c_l^{(k)} a^{(k)}. \quad (16)$$

At this point, the distribution of the test statistic can be obtained by studying the statistics of U , I and N in (16).

4.1 Interference term

Exploiting the properties of orthogonal codes, the interference term can be rewritten as

$$I = \sqrt{\frac{E_b \delta_d}{M}} \sum_{k=0, k \neq n}^{N_u-1} a^{(k)} \left(\sum_{h=1}^{\frac{M}{2}} \alpha_{x_h}^{1-\beta} - \sum_{h=1}^{\frac{M}{2}} \alpha_{y_h}^{1-\beta} \right), \quad (17)$$

where indexes x_h and y_h define the following partition

$$c^{(n)}[x_h] c^{(k)}[x_h] = 1 \quad (18)$$

$$c^{(n)}[y_h] c^{(k)}[y_h] = -1 \quad (19)$$

$$\{x_h\} \cup \{y_h\} = 0, 1, 2, \dots, M-1. \quad (20)$$

For large M , we can apply the central limit theorem (CLT) to each one of the internal sums in (17) obtaining

$$A_1, A_2 \sim \mathcal{N} \left(\sqrt{\frac{M}{2}} \mathbb{E}\{\alpha^{1-\beta}\}, \frac{M}{2} \zeta_\beta(\alpha) \right) \quad (21)$$

where $\zeta_\beta(\alpha)$ indicates the variance of $\alpha^{1-\beta}$ given by

$$\zeta_\beta(\alpha) \triangleq \mathbb{E}\{(\alpha^{1-\beta})^2\} - (\mathbb{E}\{\alpha^{1-\beta}\})^2. \quad (22)$$

Therefore, $A \triangleq A_1 - A_2$ is distributed as

$$A \sim \mathcal{N}(0, M \zeta_\beta(\alpha)). \quad (23)$$

By exploiting the symmetry of the Gaussian probability density function (p.d.f.) and the property of the sum of uncorrelated (and thus independent) Gaussian r.v.'s ($A_k = a^{(k)} A \sim (0, M \zeta_\beta(\alpha))$), the interference term results distributed as

$$I \sim \mathcal{N}(0, \sigma_I^2 \triangleq E_b \delta_d (N_u - 1) \zeta_\beta(\alpha)). \quad (24)$$

4.2 Noise term

The thermal noise at the combiner output is given by

$$N = \sum_{l=0}^{M-1} \alpha_l^{-\beta} n_l \quad (25)$$

where terms α_l and n_l are independent and n_l is zero mean. Thus, N consists on a sum of i.i.d zero mean r.v.'s with variance $N_0/2 \mathbb{E}\{\alpha^{-2\beta}\}$. By applying the CLT, we approximate the unconditioned noise term N as

$$N \sim \mathcal{N}\left(0, \sigma_N^2 \triangleq M \frac{N_0}{2} \mathbb{E}\{\alpha^{-2\beta}\}\right). \quad (26)$$

4.3 Useful term

By applying the CLT, the gain U on the useful term in (16) results distributed as

$$U \sim \mathcal{N}\left(\sqrt{E_b \delta_d M} \mathbb{E}\{\alpha_l^{1-\beta}\}, E_b \delta_d \zeta_\beta(\alpha)\right). \quad (27)$$

4.3.1 Independence between each term

By noting that $a^{(k)}$ is zero mean and statistically independent on α_l , A , and n_l , it follows that $\mathbb{E}\{I N\} = \mathbb{E}\{I U\} = 0$. Since n_l and α_l are statistically independent, the $\mathbb{E}\{N U\} = 0$. The fact that I , N and U are uncorrelated Gaussian r.v.'s implies they are also independent.

5. Bit error probability evaluation

From (24) and (26) we obtain

$$I + N \sim \mathcal{N}\left(0, E_b \delta_d (N_u - 1) \zeta_\beta(\alpha) + M \mathbb{E}\{\alpha^{-2\beta}\} \frac{N_0}{2}\right) \quad (28)$$

that can be applied to the test statistic in (16) to derive the BEP conditioned to the r.v. U as

$$P_b | u = \frac{1}{2} \operatorname{erfc}\left\{\frac{U}{\sqrt{2(\sigma_I^2 + \sigma_N^2)}}\right\}. \quad (29)$$

By applying the law of large number (LLN), that is approximating $\sum_{l=0}^{M-1} \alpha_l^{1-\beta}$ with $M \mathbb{E}\{\alpha^{1-\beta}\}$, we can derive the unconditioned BEP as

$$P_b \approx \frac{1}{2} \operatorname{erfc}\left\{\frac{\sqrt{\frac{E_b \delta_d (\mathbb{E}\{\alpha^{1-\beta}\})^2}{2 E_b \delta_d \frac{N_u - 1}{M} \zeta_\beta(\alpha) + \mathbb{E}\{\alpha^{-2\beta}\} N_0}}}{\sqrt{2 E_b \delta_d \frac{N_u - 1}{M} \zeta_\beta(\alpha) + \mathbb{E}\{\alpha^{-2\beta}\} N_0}}}\right\} \quad (30)$$

where it can be evaluated that

$$\mathbb{E}\{\alpha^{1-\beta}\} = (2\sigma_{\text{H}}^2)^{\frac{1-\beta}{2}} \Gamma\left(\frac{3-\beta}{2}\right) \quad (31)$$

$$\mathbb{E}\{\alpha^{-2\beta}\} = (2\sigma_{\text{H}}^2)^{-\beta} \Gamma(1-\beta) \quad (32)$$

$$\zeta_{\beta}(\alpha) = (2\sigma_{\text{H}}^2)^{1-\beta} \left[\Gamma(2-\beta) - \Gamma^2\left(\frac{3-\beta}{2}\right) \right] \quad (33)$$

being $\Gamma(z)$ the Euler Gamma function. Hence, we can write

$$P_{\text{b}} \approx \frac{1}{2} \operatorname{erfc} \sqrt{\frac{\Gamma^2\left(\frac{3-\beta}{2}\right) \bar{\gamma}}{2 \frac{N_{\text{u}}-1}{M} \left[\Gamma(2-\beta) - \Gamma^2\left(\frac{3-\beta}{2}\right) \right] \bar{\gamma} + \Gamma(1-\beta)}}. \quad (34)$$

where

$$\bar{\gamma} \triangleq \frac{2\sigma_{\text{H}}^2 E_{\text{b}} \delta_{\text{d}}}{N_0} \quad (35)$$

represents the mean SNR averaged over small-scale fading.

Note that the BEP expression is general in β and it is immediate to verify that results in the expressions for EGC ($\beta = 0$) and MRC ($\beta = -1$) as in (Yee et al., 1993).

As a benchmark, note also that for MRC with one active user (i.e., $N_{\text{u}} = 1$), (34) becomes

$$P_{\text{b}} \approx \frac{1}{2} \operatorname{erfc} \sqrt{\bar{\gamma}} \quad (36)$$

that is independent on the number of sub-carrier M and represents the well known limit of the antipodal waveforms in AWGN channel. This means that the approximation due to LLN is equivalent to assume that we have a number of sub-carriers (M) sufficiently high to saturate the frequency-diversity, then the transmission performs as in the absence of fading.

5.1 Optimum choice of the combining parameter

Now we will analyze the proposed PE technique with the aim of finding the optimum value of β , defined as the value within the range $[-1,1]$ that minimizes the BEP

$$\begin{aligned} \beta^{(\text{opt})} &= \arg \min_{\beta} \{P_{\text{b}}(\beta, \bar{\gamma})\} \\ &= \arg \max_{\beta} \left\{ \frac{\Gamma^2\left(\frac{3-\beta}{2}\right) \bar{\gamma}}{2 \frac{N_{\text{u}}-1}{M} \left[\Gamma(2-\beta) - \Gamma^2\left(\frac{3-\beta}{2}\right) \right] \bar{\gamma} + \Gamma(1-\beta)} \right\}. \end{aligned} \quad (37)$$

It will be shown in the numerical results that the approximation on the BEP does not significantly affect $\beta^{(\text{opt})}$. By forcing to zero the derivative of the argument in (37), after some mathematical manipulations we obtain the following expression

$$\left[\Psi\left(\frac{3-\beta}{2}\right) - \Psi(1-\beta) \right] \left[\frac{1}{\xi} + (1-\beta) \right] - 1 = 0 \quad (38)$$

where $\Psi(x)$ is the logarithmic derivative of the Gamma function, the so-called Digamma-function defined as $\Psi(x) \triangleq d\ln\Gamma(x)/dx$ (Gradshteyn & Ryzhik, 2000), and

$$\xi \triangleq 2\bar{\gamma} \frac{N_u - 1}{M} \triangleq 2\bar{\gamma} S_L \quad (39)$$

being S_L the system load. In (Zabini et al., to appear), the analysis has been extended also to derive the optimum β with imperfect channel estimation and correlated fading showing that the optimum PE parameter is not significantly affected by channel estimation errors meaning that it is possible to adopt the value of the PE parameter which would be optimum in ideal conditions even for estimation errors bigger than 1% (Zabini et al., 2007; to appear).

The parameter ξ quantifies how much the system is noise-limited (low values) or interference-limited (high values), and (38) represents the implicit solution, for the problem of finding the optimum value of β for all possible values of SNR, number of sub-carriers and number of users. Indeed, (38) open the way to an important consideration. In fact, the optimum β only depends, through ξ , on slowly varying processes such as the SNR (averaged over fast fading then randomly varying according to shadowing), the number of users and the number of sub-carriers. This means that it could be reliable an adaptive partial equalization technique in which β is slowly adapted to the optimum value for the current set of $\bar{\gamma}$, N_u and M .

6. Numerical results

In this Section, numerical results on the BEP and the optimum β in different system conditions are shown. Firstly, the goodness of the presented approach is proved by comparison with simulations. In particular, Fig. 2 shows the BEP as a function of β for different values of $\bar{\gamma}$ (5 dB and 10 dB) and $N_u = M = 1024$. Analysis and simulations appear to be in a good agreement, in particular for what concerns the value of β providing the minimum for the BEP. Moreover, it can be noted that the choice of the optimum value of β guarantees a significant improvement in the performance with respect to the cases of MRC ($\beta = -1$), EGC ($\beta = 0$) and ORC ($\beta = 1$); this improvement appears more relevant as the SNR increases.

The performance improvement of PE technique with optimum β with respect to classical MRC can be evaluated, for different system load $S_L = (N_u - 1)/M$ and SNRs, by observing Fig. 3. As an example, at $\bar{\gamma} = 8$ dB with $S_L = 20\%$ the BEP is about 0.005 with optimum β against 0.03 with MRC, whereas for $S_L = 60\%$ is about 0.015 and 0.11, for optimum β and MRC, respectively. When the system is fully-loaded, Fig. 3 also shows a comparison with MMSE (from (Slimane, 2000)) and TORC detector. For TORC we checked that $\rho_{TH} = 0.25$ is a good value for the SNR range considered. As can be observed, MMSE always provides the better performance and it is about 1 - 1.5 dB away from that obtained with PE technique with optimum β . Note also that the system with optimum β and system load 60% performs as fully-loaded MMSE.

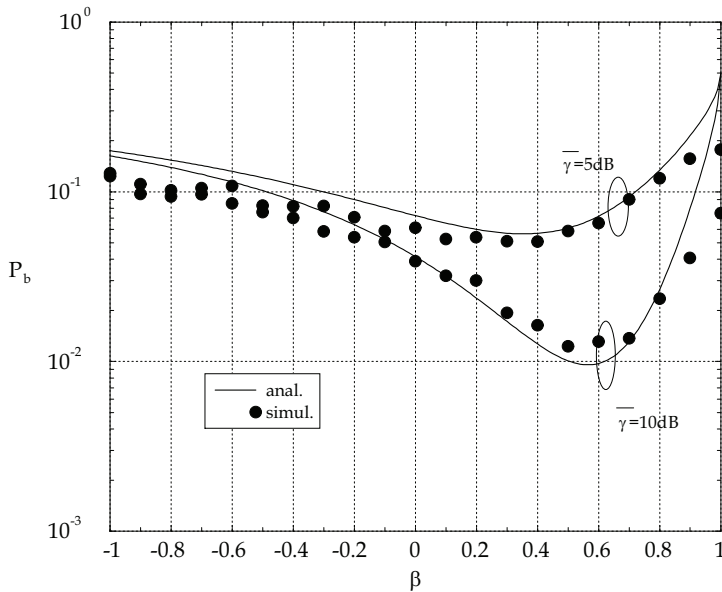


Fig. 2. BEP as a function of the PE parameter β for $\bar{\gamma} = 5$ and 10 dB in fully loaded system conditions. Comparison between analysis and simulation.

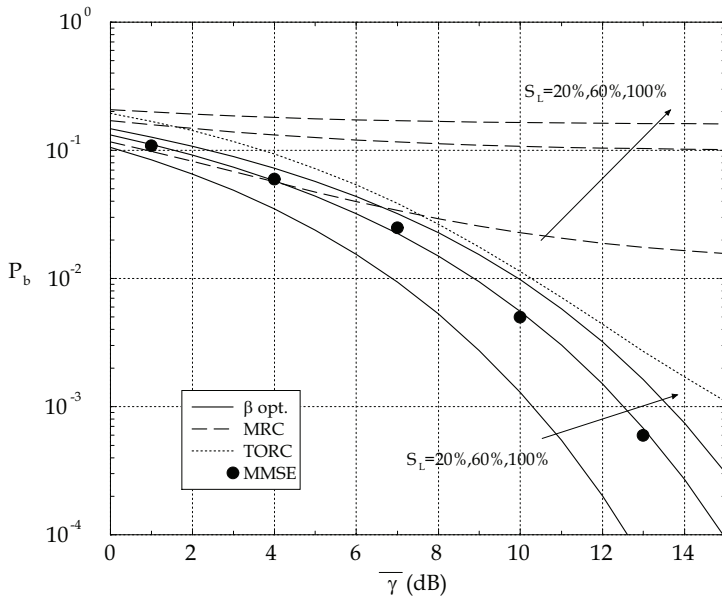


Fig. 3. BEP as a function of the mean SNR for system load $S_L = (N_u - 1) / M$ equal to 20%, 60% and fully-loaded when MRC or partial equalization with optimum β are adopted. For the fully-loaded case, the comparison includes also MMSE (from (Slimane, 2000)) and TORC detector.

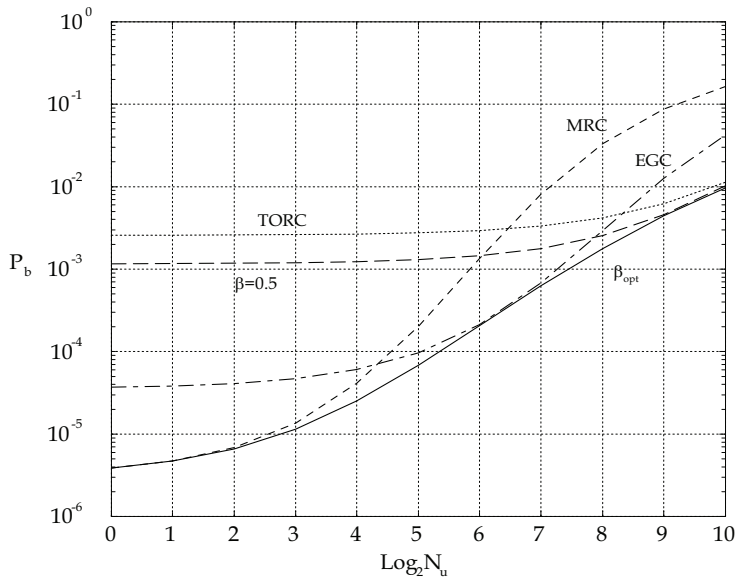


Fig. 4. The impact of the parameter β on the BEP as a function of the number of users for $M = 1024$ and $\bar{\gamma} = 10$ dB.

In Fig. 4 the impact of different equalization strategies on the BEP as a function of the number of active users, N_u , is reported for $\bar{\gamma} = 10$ dB and $M = 1024$. First of all it can be noted that the optimum β always provides the better performance; then, it can be observed that when few users are active MRC represents a good solution, approaching the optimum, crossing the performance of EGC for a system load about $1/64 \div 1/32$ (i.e., $N_u = 16 \div 32$) and the performance of a TORC detector with $\rho_{TH} = 0.25$ for a system load about $1/16 \div 1/8$. Note that a fixed value of β equal to 0.5 represents a solution close to the optimum for system loads ranging in $1/4 \div 1$ (i.e., $N_u = 256 \div 1024$) and the performance still remain in the same order for all system loads.

7. Combined equalization

Another approach to combine the sub-carriers contributions consists in applying pre-equalization at the transmitter in conjunction with post-equalization at the receiver, thereby splitting the overall equalization process on the two sides (Masini & Conti, 2009). We will call this process combined equalization (CE). The transmitter and receiver block schemes are depicted in Fig. 5.

A similar approach was proposed in (Cosovic & Kaiser, 2007), where the performance was analytically derived in the downlink for a single user case and in (Masini, 2008), where PE was considered at the transmitter and threshold ORC (TORC) at the receiver. For time division duplex direct sequence-CDMA systems a pre and post Rake receiver scheme was presented in (Barreto & Fettweis, 2000). Here we present a complete framework useful to evaluate the performance of CE (i) in a multiuser scenario; (ii) analytically evaluating optimal values for PE parameters; (iii) investigating when combined equalization introduces some benefits with respect to classical single side equalization techniques.

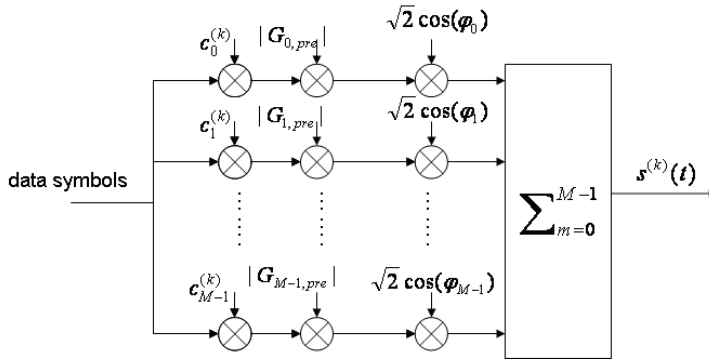
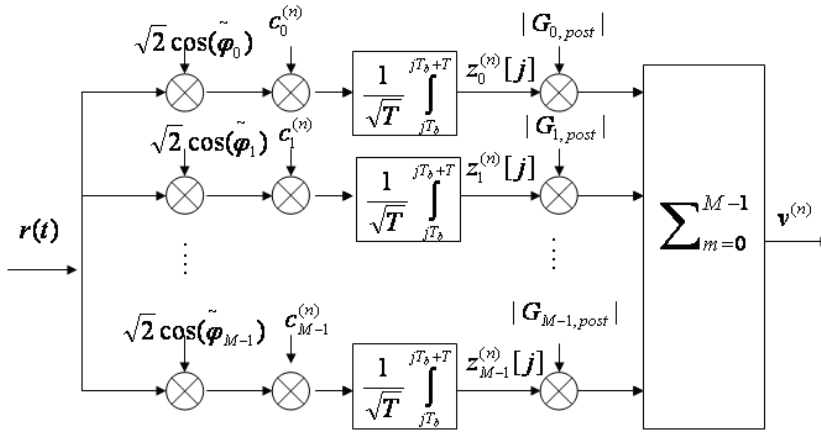
(a) Transmitter block scheme ($\varphi_m = 2\pi f_m t + \varphi_m, m = 0 \dots M-1$).(b) Receiver block scheme ($\tilde{\varphi}_m = 2\pi f_m t + \vartheta_m, m = 0 \dots M-1$).

Fig. 5. Transmitter and receiver block schemes in case of combined equalization.

We assume CSI simultaneously available at both the transmitter and the receiver in order to evaluate the impact of a combined equalization at both sides on the system performance in terms of BEP with respect to single-side equalization. In particular we assume PE performed at both sides, thus allowing the derivation of a very general analytical framework for the BEP evaluation and for the explicit derivation of the performance sensitivity to the system parameters.

7.1 Transmitter

The signal transmitted in the downlink to the totality of the users can be written as

$$s(t) = \sqrt{\frac{2E_b}{M}} \sum_{k=0}^{N_u-1} \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} c_m^{(k)} a^{(k)}[i] |G_{m,pre}| g(t - iT_b) \cos(\varphi_m). \quad (40)$$

where $G_{m,pre}$ is the pre-equalization coefficient given by

$$G_{m,\text{pre}} = G_m \sqrt{\frac{M}{\sum_{i=0}^{M-1} |G_m|^2}} \quad (41)$$

and G_m is the pre-equalization coefficient without power constraint given by (7) and here reported

$$G_m = \frac{H_m^*}{|H_m|^{1+\beta_T}} \quad (42)$$

with β_T representing the PE coefficient at the transmitter.

The coefficient $G_{m,\text{pre}}$ has to be normalized such that the transmit power is the same as in the case without pre-equalization, that means

$$\sum_{m=0}^{M-1} |G_{m,\text{pre}}|^2 = M. \quad (43)$$

Note that when $\beta_T = -1, 0$, and 1 , coefficient in (41) reduces to the case of MRC, EGC and ORC, respectively. Since we are considering the downlink we assume perfect phase compensation, the argument of $G_{m,\text{pre}}$ can be included inside ϕ_m in (40), explicitly considering only its absolute value.

Note that, to perform pre-equalization, CSI has to be available at the transmitter; this could be possible, for example, in cellular systems where the mobile unit transmits pilot symbols in the uplink which are used by the base station for channel estimation.

7.2 Receiver

By assuming the same channel model as in Sec. 3.2, the received signal results

$$r(t) = \sqrt{\frac{2E_b}{M}} \sum_{k=0}^{N_u-1} \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} \alpha_m c_m^{(k)} a^{(k)}[i] g'(t - iT_b) |G_{m,\text{pre}}| \cos(\tilde{\varphi}_m) + n(t). \quad (44)$$

At the receiver side, the post-equalization coefficient has to take into account not only the effect of channel but also of pre-equalization in order to counteract additional distortion caused by the last one. (see Fig. 5). Hence, it is given by

$$G_{l,\text{post}} = \frac{(G_{l,\text{pre}} H_l)^*}{|G_{l,\text{pre}} H_l|^{1+\beta_R}} \quad (45)$$

where β_R is the post-equalization parameter. Note again that when $\beta_R = -1, 0$ and 1 , (45) reduces to MRC, EGC and ORC, respectively.

8. Decision variable for combined equalization

Adopting the same procedure as in Sec. 4 and, hence, by linearly combining the weighted signals from each sub-carriers, we obtain the decision variable

$$v^{(n)} = \sum_{l=0}^{M-1} |G_{l,\text{post}}| z_l^{(n)} \quad (46)$$

where the received signal before combination can be evaluated as

$$z_l^{(n)}[j] = \sqrt{\frac{E_b \delta_d}{M}} \alpha_l^{1-\beta_T} \sqrt{\frac{M}{\sum_{i=0}^{M-1} \alpha_i^{-2\beta_T}}} a^{(n)}[j] + \sqrt{\frac{E_b \delta_d}{M}} c_l^{(n)} \alpha_l^{1-\beta_T} \sqrt{\frac{M}{\sum_{i=0}^{M-1} \alpha_i^{-2\beta_T}}} \sum_{k=0, k \neq n}^{N_u-1} c_l^{(k)} a^{(k)}[j] + n_l[j]. \quad (47)$$

After some mathematical manipulation

$$v^{(n)} = \overbrace{\sqrt{\frac{E_b \delta_d}{M}} \sum_{l=0}^{M-1} \alpha_l^{(1-\beta_T)(1-\beta_R)} a^{(n)}}^U + \overbrace{\sqrt{\frac{E_b \delta_d}{M}} \sum_{l=0}^{M-1} \sum_{k=0, k \neq n}^{N_u-1} \alpha_l^{(1-\beta_T)(1-\beta_R)} c_l^{(n)} c_l^{(k)} a^{(k)}}^I + \underbrace{\sum_{l=0}^{M-1} \alpha_l^{-\beta_R(1-\beta_T)} n_l \sqrt{\frac{\sum_{i=0}^{M-1} \alpha_i^{-2\beta_T}}{M}}}_N \quad (48)$$

where U , I , and N represent the useful, interference, and noise term, respectively and whose statistic distribution has to be derived to evaluate the BEP.

Following the same procedure adopted in Sec. 4, we obtain

$$U \sim \mathcal{N}\left(\sqrt{E_b \delta_d M} \mathbb{E}\{\alpha_l^{(1-\beta_T)(1-\beta_R)}\}, \sigma_U^2\right) \quad (49)$$

$$I \sim \mathcal{N}\left(0, \sigma_I^2 = E_b \delta_d (N_u - 1) (2\sigma_H^2)^{(\beta_T-1)(\beta_T-1)}\right) \quad (50)$$

$$\times \left(\Gamma[2 + \beta_T(\beta_R - 1) - \beta_R] - \Gamma^2 \left[\frac{3 + \beta_T(\beta_R - 1) - \beta_R}{2} \right] \right) \quad (51)$$

$$N \sim \mathcal{N}\left(0, \sigma_N^2 = M \frac{N_0}{2} (2\sigma_H^2)^{-\beta_T + \beta_R(\beta_T-1)} \Gamma[1 - \beta_T] \Gamma[1 + \beta_R(\beta_T - 1)]\right). \quad (52)$$

Also in this case, since $a^{(k)}$ is zero mean and statistically independent of α_l and n_l , and considering that n_l and α_l are statistically independent and zero mean too, then $\mathbb{E}\{IN\} = \mathbb{E}\{IU\} = 0$. Since n_l and α_l are statistically independent, then $\mathbb{E}\{NU\} = 0$. Moreover I , N , and U are uncorrelated Gaussian r.v.'s, thus also statistically independent.

9. Bit error probability evaluation with combined equalization

By applying the LLN to the useful term, that is by approximating U with its mean value, the BEP averaged over small-scale fading results

$$P_b \approx \frac{1}{2} \operatorname{erfc} \sqrt{\Xi}, \quad (53)$$

where Ξ is the signal-to-noise plus interference-ratio (SNIR) given by

$$\Xi \triangleq \frac{\bar{\gamma} \Gamma^2 \left[\frac{3+\beta_T(\beta_R-1)-\beta_R}{2} \right]}{\Gamma[1-\beta_T]\Gamma[1+\beta_R(\beta_T-1)] + 2\bar{\gamma} \frac{N_u-1}{M} \left(\Gamma[2+\beta_T(\beta_R-1)-\beta_R] - \Gamma^2 \left[\frac{3+\beta_T(\beta_R-1)-\beta_R}{2} \right] \right)} \quad (54)$$

Note that when one between β_T or β_R is zero, (53) reduces to (34).

10. Optimum combination with combined equalization

We aim at deriving the optimal choice of the PE parameters, thus the couple (β_T, β_R) jointly minimizing the BEP

$$(\beta_T, \beta_R)^{(\text{opt})} = \arg \min_{\beta_T, \beta_R} \{P_b(\beta_T, \beta_R, \bar{\gamma})\}. \quad (13)$$

However, being in the downlink, the receiver is in the mobile unit, hence, it is typically more convenient, if necessary, to optimize the combination at the transmitter (i.e., at the base station), once fixed the receiver. Therefore, we find the optimum values of β_T defined as that values within the range $[-1,1]$ that minimizes the BEP for each β_R

$$\beta_T^{(\text{opt})} = \arg \min_{\beta_T} \{P_b(\beta_T, \beta_R, \bar{\gamma})\} \approx \arg \max_{\beta_T} \{\Xi\}. \quad (14)$$

By deriving (54) with respect to β_T and after some mathematical manipulation, we obtain the implicit solution given by (15)

$$\xi = \frac{\Gamma[1-\beta_T]\Gamma[1+\beta_R(\beta_T-1)]}{(\beta_R-1)\Gamma[2+\beta_T(\beta_R-1)-\beta_R] \left\{ \Psi \left[\frac{3+\beta_T(\beta_R-1)-\beta_R}{2} \right] - \Psi[2+\beta_T(\beta_R-1)-\beta_R] \right\}} \times \left\{ -(\beta_R-1)\Psi \left[\frac{3+\beta_T(\beta_R-1)-\beta_R}{2} \right] - \Psi[1-\beta_T] + \beta_R \Psi[1+\beta_R(\beta_T-1)] \right\}. \quad (15)$$

11. Numerical results for combined equalization

In Fig. 6, the BEP is plotted as a function of β_T for different values of β_R and mean SNR $\bar{\gamma} = 10$ dB in fully loaded system conditions ($M = N_u = 1024$). Note that, in spite of the post-PE technique, there is always an optimum value of β_T minimizing the BEP and this value depends on β_R . Moreover, the BEP is also drastically dependent on β_R , meaning that a not suitable post-PE technique can even deteriorate the performance, with respect to one side combination, rather than improving it. Simulation results are also reported confirming the analysis especially in correspondence to the optimal β_R (note that the analysis is confirmed for 64 sub-carriers and thus it is expected to be even more accurate for higher number of sub-carriers).⁵

⁵ Similar considerations can be drawn for time- and -frequency correlated SUI-x channels as shown, by simulation, in (Masini et al., 2008) referred to PE at the receiver.

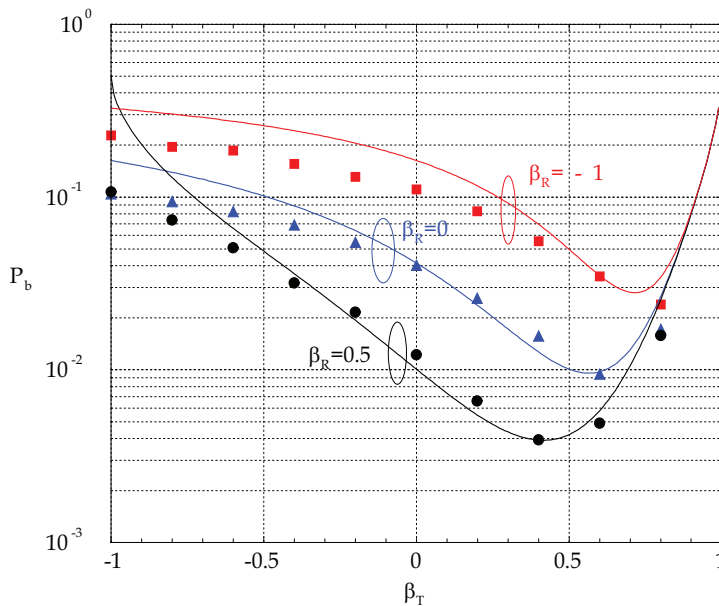


Fig. 6. BEP vs. the pre-equalization parameter β_T for different post-equalization parameter values β_R and $\bar{\gamma} = 10$ dB in fully loaded system conditions. Comparison between analysis and simulation. Figure reprinted with permission from B. M. Masini, A. Conti, "Combined Partial Equalization for MC-CDMA Wireless Systems", IEEE Communications Letters, Volume 13, Issue 12, December 2009 Page(s):884 – 886. ©2009 IEEE.

In Fig. 7, the BEP is plotted as a function of the mean SNR, $\bar{\gamma}$, in fully loaded system conditions ($M = N_u = 1024$). The effect of the combining techniques at the transmitter and the receiver can be observed: a suitable choice of coefficients (such as $\beta_T = 0.5$ and $\beta_R = 0.5$) improves the performance with respect to single side combination ($\beta_T = 0$, $\beta_R = 0.5$); however, a wrong choice (such as $\beta_T = 0.5$ and $\beta_R = -1$) can drastically deteriorate the BEP.

In Fig. 8, the BEP as a function of the system load S_L in percentage is shown for $\bar{\gamma} = 10$ dB and different couples (β_T , β_R). Note how a suitable choice of pre- and post-PE parameters can increase the sustainable system load. At instance, by fixing a target BEP equal to $4 \cdot 10^{-3}$, with combination at the transmitter only (i.e., $\beta_T = 0.5$, $\beta_R = 0$) we can serve the 45% of users, while fixing $\beta_T = 0.5$ and adaptively changing β_R following the system variations (i.e., always setting β_R at the optimum value minimizing the BEP), the 100% of users can be served. The same performance can be obtained by fixing the combination parameter at 0.5 at the transmitter or at the receiver and adaptively changing the combination parameter at the

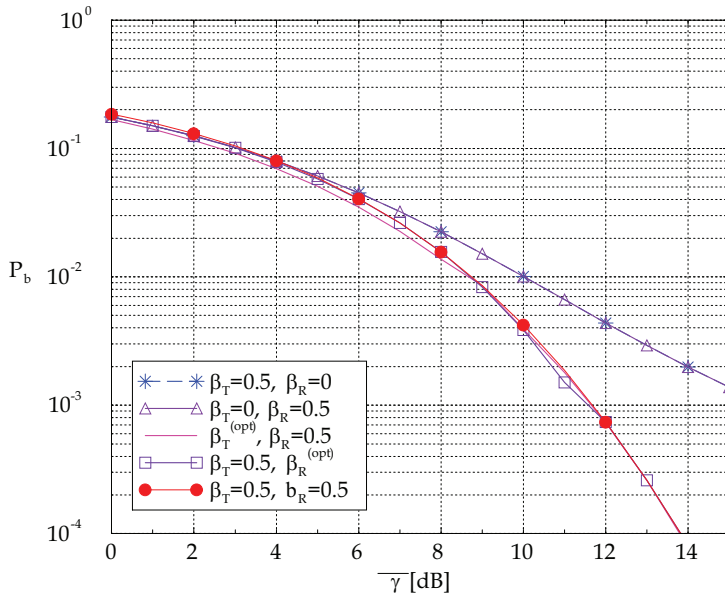


Fig. 7. BEP vs. the mean SNR $\bar{\gamma}$ for different couples of β_T and β_R in fully loaded system conditions.

other side. The same performance can also be obtained by exploiting the couple of fixed parameter ($\beta_T = 0.5, \beta_R = 0.5$), thus avoiding the complexity given by parameters adaptation. It is also worth noting that a not suitable choice of combination parameters, such as ($\beta_T = -0.5, \beta_R = 0$) or ($\beta_T = 0.5, \beta_R = -0.5$) can even deteriorate the performance with respect to single side combination.

12. Final considerations

We summarized the main characteristics of MC-CDMA systems and presented a general framework for the analytical performance evaluation of the downlink of MC-CDMA systems with PE.

We can conclude that MC-CDMA systems may be considered for next generation mobile radio systems for their high spectral efficiency and the low receiver complexity due to the avoidance of ISI and ICI in the detection process. The spreading code length can be dynamically changed and not necessarily equal to the number of sub-carriers enabling a flexible system design and further reducing the receiver complexity.

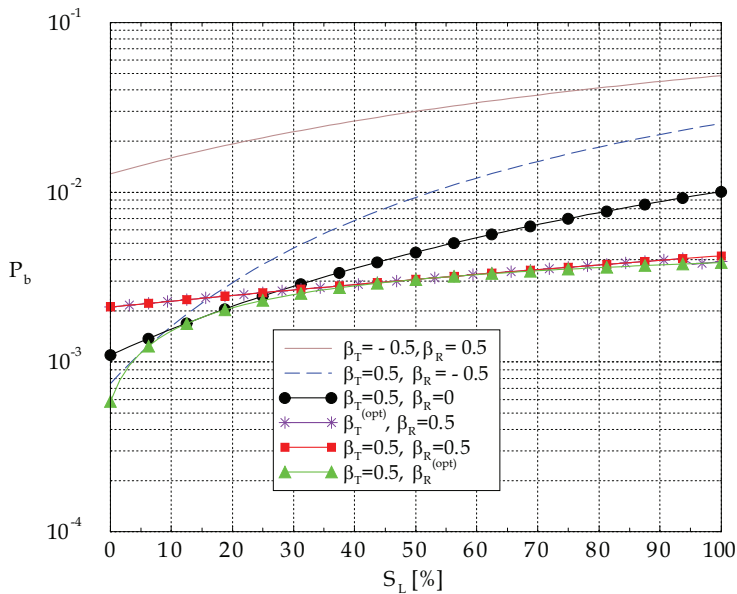


Fig. 8. BEP vs. the system load S_L for various β_T and β_R when $\bar{\gamma} = 10$ dB. Figure reprinted with permission from B. M. Masini, A. Conti, "Combined Partial Equalization for MC-CDMA Wireless Systems", IEEE Communications Letters, Volume 13, Issue 12, December 2009 Page(s):884 - 886. ©2009 IEEE.

To enhance their performance, PE can be adopted in the downlink, allowing good performance in fading channels still maintaining low the receiver complexity.

The optimal choice of the PE parameter is fundamental to improve the performance in terms of BEP averaged over small-scale fading.

When CE is adopted at both the transmitter and the receiver a proper choice of PE parameters is still more important, to significantly improve the performance with respect to single-side detection.

The gain achieved by a suitable combination of transmission and reception equalization parameters could be exploited to save energy or increase the coverage range (a similar approach was used for partial power control in cellular systems in (Chiani et al., 2001)).

In case of non-ideal channel estimation, the performance results to be deteriorated; however, it has been shown that the optimum PE parameter is not significantly affected by channel estimation errors. The analysis for correlated fading channels and imperfect CSI has been

performed in (Zabini et al., to appear). optimum PE parameter with perfect CSI This means that, in practical systems, it is possible to adopt the value of the PE parameter which would be optimum in ideal conditions (it is simple to evaluate and does not require the knowledge of the channel estimation error) without a significant loss of performance, even for estimation errors bigger than 1% (Zabini et al., 2007; to appear).

The effect of block fading channels and time and frequency correlated fading channel on the performance of MC-CDMA systems with PE has been investigated in (Masini & Zabini, 2009) and (Masini et al., 2008), respectively, still showing the goodness of PE as linear equalization technique and still demonstrating that the PE parameter that is optimum in ideal scenarios still represents the best choice also in more realistic conditions.

13. References

- Barreto, A. & Fettweis, G. (2000). Performance improvement in ds-spread spectrum cdma systems using a pre- and a post-rake, Zurich, pp. 39–46.
- Chiani, M., Conti, A. & Verdone, R. (2001). Partial compensation signal-level-based up-link power control to extend terminal battery duration, *Vehicular Technology, IEEE Transactions on* 50(4): 1125–1131.
- Conti, A., Masini, B., Zabini, F. & Andrisano, O. (2007). On the down-link performance of multi-carrier CDMA systems with partial equalization, *IEEE Transactions on Wireless Communications* 6(1): 230–239.
- Cosovic, I. & Kaiser, S. (2007). A unified analysis of diversity exploitation in multicarrier cdma, *IEEE Transactions on Vehicular Technology* 56(4): 2051–2062.
- Gradshteyn, I. & Ryzhik, I. (2000). *Table of Integrals, Series and Products*, Academic Press.
- Hanzo, L. & Keller, T. (2006). *OFDM and MC-CDMA - A Primer*, J. Wiley & Sons. ISBN: 0470030070.
- Hanzo, L., Yang, L.-L., Kuan, E.-L. & Yen, K. (2003). *Single and Multi-Carrier DS-CDMA: Multi-User Detection, Space-Time Spreading, Synchronization and Standards*, J.Wiley & Sons.
- K. Fazel, S. K. (2003). *Multi-Carrier and Spread Spectrum Systems*, Wiley.
- Masini, B. (2008). The impact of combined equalization on the performance of mc-cdma systems, *Journal of Communications* 3(5): 2051–2062.
- Masini, B. & Conti, A. (2009). Combined partial equalization for MC-CDMA wireless systems, *IEEE Communications Letters* 13(12): 884–886.
- Masini, B., Leonardi, G., Conti, A., Pasolini, G., Bazzi, A., Dardari, D. & Andrisano, O. (2008). How equalization techniques affect the tcp performance of mc-cdma systems in correlated fading channels, *EURASIP Journal on Wireless Communications and Networking* (Article ID 286351).
- Masini, B. & Zabini, F. (2009). On the effect of combined equalization for mc-cdma systems in correlated fading channels, *IEEE Wireless Communications and Networking Conference, WCNC*, pp. 1–6.
- Slimane, S. (2000). Partial equalization of multi-carrier cdma in frequency selective fading channels, New Orleans, USA, pp. 26–30.

- Yee, N., Linnartz, J.-P. & Fettweis, G. (1993). Multi-Carrier-CDMA in indoor wireless networks, *Proceedings of Personal, Indoor and Mobile Radio Conference, PIMRC*, Yokohama, pp. 109-113.
- Zabini, F., Masini, B. & Conti, A. (2007). On the performance of MC-CDMA systems with partial equalization in the presence of channel estimation errors, *6th IEEE International Workshop on Multi Carrier Spread Spectrum (MC-SS)*, Herrsching, Germany, pp. 407- 416.
- Zabini, F., Masini, B., Conti, A. & Hanzo, L. (to appear). Partial equalization for MC-CDMA systems in non-ideally estimated correlated fading, *IEEE Transactions on Vehicular Technology*.

Wireless Multimedia Communications and Networking Based on JPEG 2000

Max AGUEH
ECE Paris
France

1. Introduction

Nowadays, more and more multimedia applications integrate wireless transmission functionalities. Wireless networks are suitable for those types of applications, due to their ease of deployment and because they yield tremendous advantages in terms of mobility of User Equipment (UE). However, wireless networks are subject to a high level of transmission errors because they rely on radio waves whose characteristics are highly dependent of the transmission environment.

In wireless video transmission applications like the one considered in this chapter and presented in Figure 1, effective data protection is a crucial issue.

JPEG 2000, the newest image representation standard, addresses this issue firstly by including predefined error resilient tools in his core encoding system (part 1) and going straightforward by defining in its 11th part called wireless JPEG 2000 (JPWL) a set of error resilient techniques to improve the transmission of JPEG 2000 codestreams over error-prone wireless channel.

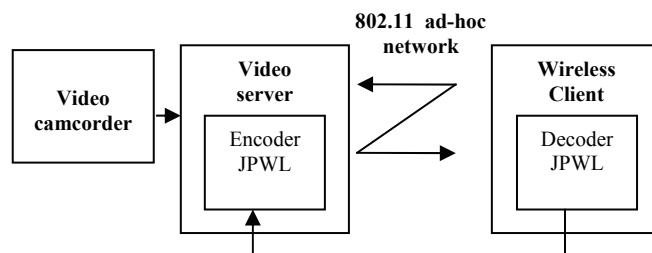


Fig. 1. Wireless video streaming system

JPEG 2000

JPEG 2000 is the newest image compression standard completing the existing JPEG standard (Taubman & Marcellin, 2001).

The interest for JPEG 2000 is growing since the Digital Cinema Initiatives (DCI) has selected JPEG 2000 for future distribution of motion pictures.

Its main characteristics are: lossy or lossless compression modes; resolution, quality and spatial scalability; transmission and progressive image reconstruction; error resilience for low bit rate mobile applications; Region Of Interest (ROI) functionality, etc.

Part 1 of the standard defines different tools allowing the decoder to detect errors in the transmitted codestream, to select the erroneous part of the code and to synchronise the decoder in order to avoid decoder crash. Even if those tools give a certain level of protection from transmission errors, they become ineffective when the transmission channel experiment high bit error rate. Wireless JPEG 2000 (JPEG 2000 11th part) addressed this issue by defining techniques to make JPEG 2000 codestream more resilient to transmissions errors in wireless systems.

Wireless JPEG 2000 (JPWL)

Wireless JPEG (JPWL) specifies error resilience tools such as Forward Error correction (FEC), interleaving, unequal error protection.

In this chapter we present a wireless JPEG 2000 video streaming system based on the recommendations of JPWL final draft (JPWL, 2005).

In (Dufaux & Nicholson, 2004), the description of the JPWL system is presented and the performance of its Error Protection Block (EPB) is evaluated. A fully JPEG 2000 Part 1 compliant backward compatible error protection scheme is proposed in (Nicholson et al, 2003). A memoryless Binary Symmetric Channel (BSC) is used for simulations both in (Nicholson et al, 2003) and (Dufaux & Nicholson, 2004). However, as packets errors mainly occur in bursts, the channel model considered in those works is not realistic. Moreover JPEG 2000 codestreams interleaving is not considered in (Nicholson et al, 2003).

In this chapter we address the problem of robust and efficient JPEG 2000 images and video transmission over wireless networks. The chapter is organized as follows: In section 2, we present a state of art of wireless JPEG 2000 multimedia communication systems along with the challenges to overcome in terms of codestreams protection against transmission errors. In section 3, we provide an overview of channel coding techniques for efficient JPEG 2000 based multimedia networking. Finally section 4, provides discussions and prospective issues for future distribution of motion JPEG 2000 images and video over wireless networks.

2. Wireless JPEG 2000 multimedia communication system and its challenges

In high error rate environments such as wireless channels, data protection is mandatory for efficient transmission of images and video. In this context, Wireless JPEG 2000 (JPWL) the 11th part of JPEG 2000 (JPWL, 2005) different techniques such as data interleaving, Forward Error Correction (FEC) with Reed-Solomon (RS) codes etc. in order to enhance the protection of JPEG 2000 codestreams against transmission errors.

In wireless multimedia system such as the one considered in this chapter (see Figure 1), a straightforward FEC methodology is applying FEC uniformly over the entire stream (Equal Error Correction - EEP). However, for hierarchical codes such as JPEG 2000, Unequal Error Protection (UEP) which assigns different FEC to different portion of codestream has been considered as a suitable protection scheme.

Since wireless channels' characteristics depend on the transmission environment, the packet loss rate in the system also changes dynamically. Thus a priori FEC rate allocation schemes such as the one proposed in (Agueh et al, 2007, a) are less efficient. Two families of data protection schemes address this issue by taking the wireless channel characteristics into

account in order to dynamically assign the FEC rate for JPEG 2000 based images/video. The first family is based on a dynamic layer-oriented unequal error protection methodology whereas the second relies on a dynamic packet-oriented unequal error protection methodology. Hence, in the first case, powerful RS codes are assigned to most important layers and less robust codes are used for the protection of less important layers. It is worth noting that in this case, all the JPEG 2000 packets belonging to the same layer are protected with the same selected RS code. Examples of layer-oriented FEC rate allocation schemes are available in (Guo et al, 2006) and (Agueh et al, 2007, b). On the other side, in packet-oriented FEC rate allocation schemes such as the one presented in (Agueh et al, 2008), RS codes are assigned by decreasing order of packets importance. In (Agueh et al, 2008), we demonstrate that the proposed optimal packet-oriented FEC rate allocation is more efficient than the layer-oriented FEC rate allocation scheme presented in (Guo et al, 2006) and (Agueh et al, 2007, b). However, layer-based FEC rate allocation schemes have low complexity while packet-oriented FEC allocation methodologies are complex especially when the number of packets in the codestream is high. In this case, packet oriented FEC schemes are unpractical for highly time-constrained images/video streaming applications. In this case switching to a layer oriented FEC rate allocation scheme is more interesting. The smart FEC rate allocation scheme proposed in (Agueh et al, 2009, a) address this issue by allowing switching from a packet oriented FEC scheme to a layer oriented scheme such as the ones proposed in (Agueh et al, 2009, b).

In section 2.1 we present the packet oriented system proposed in (Agueh et al, 2008) to address the issue of robust JPEG 2000 images and video transmission over wireless network. Then in section 2.2 the layer-oriented scheme proposed in (Agueh et al, 2009, b) is described. Finally, in section 2.3 we present the system proposed in (Agueh et al, 2009, a) to unify packet and layer based scheme.

2.1 Optimal Packet-oriented FEC rate allocation scheme for robust Wireless JPEG 2000 based multimedia transmission

The functionalities of the proposed JPWL packet-oriented system are presented in Figure 2. The aim of this system is to efficiently transmit a Motion JPEG 2000 (MJ2) video sequence through MANET channel traces.

The system is described as follows:

The input of the JPWL codec is a Motion JPEG 2000 (MJ2) file. The JPEG 2000 codestreams included in the MJ2 file are extracted and indexed.

These indexed codestreams are transmitted to the JPWL encoder (JPWL, 2005) presents a more accurate description of the used JPWL encoder) which applies FEC at the specified rate and adds the JPWL markers in order to make the codestream compliant to Wireless JPEG 2000 standard. At this stage, frames are still JPEG 2000 part 1 compliant, which means that any JPEG 2000 decoder is able to decode them.

To increase JPWL frames robustness, an interleaving mechanism is processed before each frame transmission through the error-prone channel. This is a recommended mechanism for transmission over wireless channel where errors occur in burst (contiguous long sequence of errors). Thanks to interleaving the correlation between error sequences is reduced.

The interleaving step is followed by RTP packetization. In this process, JPEG 2000 codestream data and other types of data are integrated into RTP packets as described in (Schulzrinne et al, 2003).

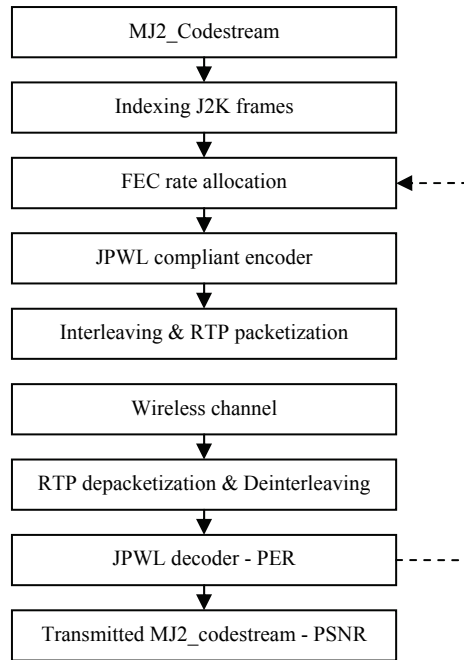


Fig. 2. JPWL based system functionalities

RTP packets are then transmitted through the wireless channel which is modelled in this work by a Gilbert channel model. At the decoder side, after depacketization, the JPWL decoder corrects and decodes the received JPWL codestreams and rebuilds the JPEG 2000 frames. At this stage, parameters such as Packet Error Rate (PER) are extracted, increasing the knowledge of the channel state. The decoder sends extracted parameters back to the JPWL encoder via the Up link. The last process of the transmission chain is the comparison between the transmitted and the decoded image/video. Figure 3 presents JPEG 2000 codestreams transmission through the JPWL packet-oriented FEC system

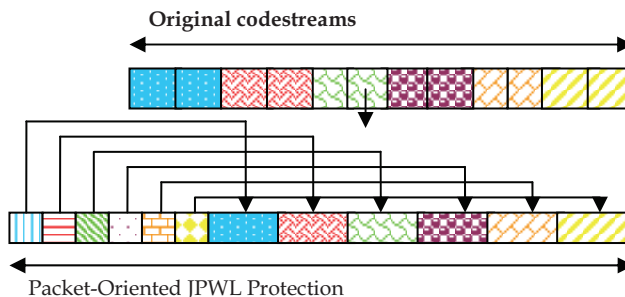


Fig. 3. JPEG 2000 codestreams transmission through the JPWL packet-oriented FEC system

2.2 Optimal Layer-oriented FEC rate allocation scheme for robust Wireless JPEG 2000 based multimedia transmission

Unlike the system described in (Agueh et al, 2008), where the FEC rate allocation scheme is packet oriented, in the current system we consider a layer oriented FEC rate allocation scheme. In other words the difference between both systems is the FEC rate allocation module. Actually, in the packet oriented scheme the redundancy is added by taking the packets importance into account (see Figure 3) while in the layer oriented scheme we rely on layers importance to allocate the adequate RS codes (see Figure 4).

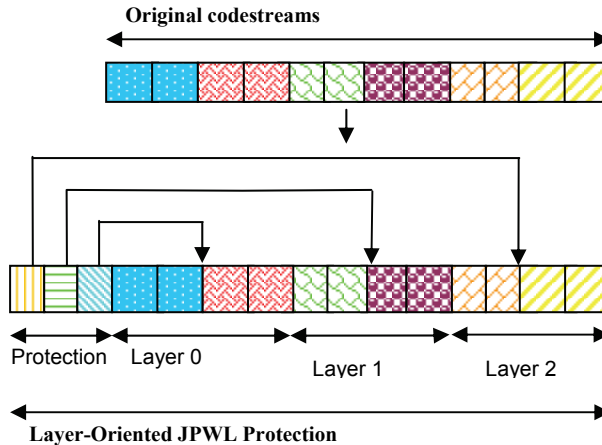


Fig. 4. A JPEG 2000 codestreams transmission through the JPWL layer-oriented FEC system

2.3 Smart combined Packet/layer based FEC rate allocation scheme for robust Wireless JPEG 2000 based multimedia transmission

The functionalities of the proposed smart JPWL based system are presented in Figure 5. In this system, indexed JPEG 2000 codestreams are transmitted to the smart FEC rate allocation module. If the number of data packets available in the codestreams is low (typically under the defined smart threshold), the smart module uses the optimal packet-oriented FEC rate allocation methodology presented in (Agueh et al, 2008) whereas it switches to the dynamic layer-oriented FEC rate allocation methodology presented in (Agueh et al, 2009, b) when the number of data packets is high. Once the protection rate is determined, the codestreams are transmitted to the JPWL encoder which applies FEC at the specified rate and adds the JPWL markers in order to make the codestream compliant to the Wireless JPEG 2000 standard. Hence, Figures 3 and 4 correspond to the JPWL protection where redundant data are added to original codestreams. If the JPEG 2000 Frame which is being processed is constituted by less than a defined threshold ($smart_thresh$), then the smart FEC rate allocation scheme emulates a scenario similar to the one presented in Figure 3 (packet-oriented FEC rate allocation). Otherwise, it emulates the scenario of Figure 4 (dynamic layer-oriented FEC rate allocation). Protected data are then interleaved and the interleaved codestreams go through the other processes described in section 2.1.

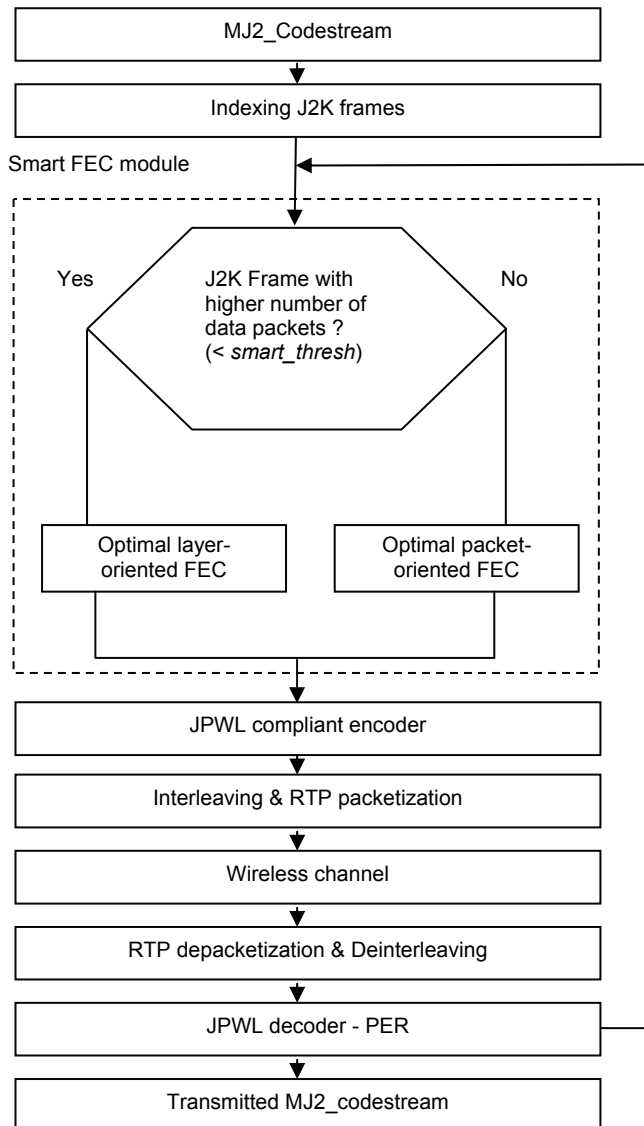


Fig. 5. JPEG 2000 transmission over the smart JPWL system

The interest of the smart FEC rate allocation scheme is to allow switching from the scenario presented in Figure 3 to the scenario described in Figure 4, reducing by this way the complexity of the FEC rate allocation process. Hence, in case of highly layered images/video streaming, the time needed to select the suited FEC rate is significantly reduced. In the following section we formalised the problem of FEC rate allocation, then we present the packet-oriented and layer-oriented algorithm considered in this book.

3. Channel coding techniques for robust wireless JPEG 2000 networking

3.1 Optimal packet-oriented FEC rate allocation for robust JPEG 2000 image and video transmission over wireless networks

Making an analogy between the FEC rate allocation problem and the Multiple-Choice Knapsack Problem (MCKP) leads to the conclusion that both problems are NP-hard. Hence, most of the algorithms proposed in the literature such as the one presented in (Thomos et al, 2004), lead to exhaustive search among different FEC rate solutions, exponentially increasing their complexity. These algorithms are thus interesting for an offline video streaming but are unpractical for real-time applications.

To overcome this limitation, Z. Guo et al proposed in (Guo et al, 2006) a slightly complex layered unequal error protection scheme for robust Motion JPEG 2000 streaming over wireless network. However, this algorithm is not JPWL compliant and was designed based on the assumption that the channel is a memoryless Binary Symmetric Channel (uncorrelated error occurrence) which is not realistic because wireless channels have correlated errors sequence. Hence, we have proposed in (Agueh et al, 2007,b) a dynamic layered based unequal error protection FEC rate allocation methodology for efficient JPEG 2000 streaming over MANET. The proposed scheme improved the performance by about 10% compared to a priori selection of channel coding (Agueh et al, 2007,a). However the main drawback of both methodologies is that the FEC rate allocation is suboptimal. In fact, in both schemes the protection strategy is layer based which implies that a selected FEC rate is applied to all the substreams belonging to the same layer. This limits the effectiveness of those protection strategies especially for fast varying channels where the selected FEC rate may need to be updated from one substream to another.

In (Agueh et al, 2008) we propose a slightly complex, packet based optimal FEC rate allocation algorithm for robust Motion JPEG 2000 video streaming over wireless channel.

3.1.a Problem formalization

The goal is to optimally protect JPEG 2000 images/video for robust streaming over wireless channel.

Considering that JPEG 2000 codestreams are constituted by a set of S substreams, the optimal FEC allocation problem can be resumed by answering the question: How to optimally protect each substream so as to minimize the transmitted image distortion under a rate constraint determined by the available bandwidth in the system?

Since the JPEG 2000 standard specifies that packets are byte aligned, it is especially interesting to work with Galois Field $GF(2^8)$ to provide error correction capabilities. In this context, JPWL final draft (JPWL, 2005) recommends the use of Reed-Solomon (RS) codes as FEC codes and fixes a set of RS default codes for substream protection before transmission over wireless channels.

Let γ be a substream protection level selected in the range $0 \leq \gamma \leq \gamma_{\max}$, each protection level corresponds to a specific RS code selected between JPWL default RS codes ($\gamma=0$ means that the substream is not transmitted, $\gamma=1$ means transmission with protection level 1, higher values imply increasing channel code capacity with γ).

Let B_{av} the byte budget constraint corresponding to the available bandwidth in the system.

Let l_i the length in bytes of the i^{th} packet of the S substreams and $RS(n,k)$ the Reed-Solomon code used for its protection, the corresponding protection level is γ and the FEC

coding rate is $R = \frac{k}{n}$. We define $fec = \frac{1}{R_i} = \frac{n}{k}$ as the invert of the channel coding rate, so $l_i * fec$ represents, in byte, the increase of the i^{th} packet length when protected at level γ . The correct decoding of packet i at the receiver yields a reduction of the distortion on the transmitted image. Let RD_i^0 be respectively the reduction of distortion associated to decoding of packet i and $RD_{i,\gamma}$ the reduction of distortion achieved when packet i is protected at level γ . The reduction of distortion metric associated to the correct decoding of the packets of a JPEG 2000 codestream is extracted from a codestream index file. The codestream index file is generated by the OpenJPEG library (<http://www.openjpeg.org>) and defines the gain in quality and the range of bytes corresponding to each packet. The reduction of distortion metric is presented in (Agueh et al, 2008).

We define the gain as the ratio between the image quality improvement $RD_{i,\gamma}$ and the associated cost in terms of bandwidth consumption $l_i * fec$.

Thus, the FEC allocation problem becomes: How to optimally select substream i protection level γ so as to maximize the associated reduction of distortion $RD_{i,\gamma}$ under a budget constraint B_{av} . This problem is formalised by:

$$\text{Maximize } \sum_{i=1}^S \frac{RD_{i,\gamma}}{l_i \cdot fec_i} \quad (1)$$

$$\text{Subject to } \sum_{i=1}^S l_i \cdot fec_i \leq B_{av} \quad (2)$$

3.1.b Optimization

Since the optimization problem can be solved by finding the optimal protection for each substream of JPEG 2000 codestreams under a budget constraint, we define $G_{i,\gamma}$ as the gain in quality of the transmitted image obtained at the receiver side when packet i is decoded.

Let $RD_{i,1}$ and $RD_{i,\gamma}$ be the reduction of distortion obtained when packet i is transmitted respectively with protection level 1 and with protection level γ , we have:

$$RD_{i,1} = (1 - P_{pack}^{i,1}) \cdot RD_{pack}^i \quad \text{and} \quad RD_{i,\gamma} = (1 - P_{pack}^{i,\gamma}) \cdot RD_{pack}^i \quad (3)$$

Where $P_{pack}^{1,\gamma}$ and $P_{pack}^{i,\gamma}$ are the decoding error probabilities obtained when packet i is protected respectively to level 1 and to level γ . The resulting gain is:

$$G_{i,1} = \frac{RD_{i,1}}{l_i} = \frac{(1 - P_{pack}^{i,1}) \cdot RD_{pack}^i}{l_i} \quad (4)$$

Similarly, any transmission between two consecutive protection levels (γ and $\gamma + 1$) yields an improvement in terms of reduction of distortion but has a budget cost equal to $(fec_{\gamma+1} - fec_{\gamma}) \times l_i$, hence we have:

$$G_{i,\gamma} = \frac{RD_{i,\gamma} - RD_{i,\gamma-1}}{(fec_{\gamma} - fec_{\gamma-1}) \cdot l_i} \quad (5)$$

$$G_{i,\gamma} = \frac{(P_{pack}^{i,\gamma-1} - P_{pack}^{i,\gamma}) \cdot RD_{pack}^i}{(fec_{\gamma} - fec_{\gamma-1}) \cdot l_i} \quad (6)$$

Protection levels incremental gains $G_{1,1}$ to $G_{S,\gamma}$ are derived for each packet and stored in S different vectors. After merging and reorganizing those vectors, the optimal protection level is derived from the maximum related gain value selected when meeting the rate constraint (Bandwidth available B_{av}). A detailed description of the processes is available in (Agueh et al, 2008).

3.1.c Synopsis of the FEC rate allocation scheme and algorithm

Synopsis of the optimal FEC rate allocation algorithm:

Algorithm:

For each JPEG 2000 image

- Model the channel with a Gilbert model and for each possible protection level γ , evaluate the probability of incorrect word decoding $P_{pack}^{i,\gamma}$
- For $i = 1$ to $i = S$ (Number of JPEG 2000 packets)

For $\gamma = 1$ to $\gamma = \gamma_{max}$

$$\text{Estimate } RD_{i,\gamma} = (1 - P_{pack}^{i,\gamma}) \cdot RD_{pack}^i$$

$$G_{i,\gamma} = \frac{RD_{i,\gamma} - RD_{i,\gamma-1}}{(fec_{\gamma} - fec_{\gamma-1}) \cdot l_i}$$

$$V(i)[\gamma] = G_{i,\gamma}$$

End For

- Merging $V(i)$ vectors protection levels if necessary to ensure that $V(i)$ vectors are constituted of strictly decreasing gains values
- Collecting V_{all}

End For

- Ordering V_{all} on decreasing order of importance values ($V_{all_ordered}$)
- Selecting each gain value, corresponding to a specific protection level, up to meeting the rate constraint
- Optimally protect JPEG 2000 packets with the corresponding Reed-Solomon codes

End For

3.2 Optimal Layer-oriented FEC rate allocation for robust JPEG 2000 image and video transmission over wireless networks

3.2.a Problem formalization

Considering that JPEG 2000 codestreams are constituted by a set of L layers, the optimal FEC allocation problem can be resumed by answering the question: How to optimally protect each layer in order to minimize the transmitted image distortion under a rate constraint determined by the available bandwidth in the system?

Let lay_i the length in bytes of the i^{th} layer of the L layers and $RS(n,k)$ the Reed-Solomon code used for its protection, the corresponding protection level is γ and the FEC coding rate is $R = \frac{k}{n}$.

We define $fec = \frac{1}{R} = \frac{n}{k}$ as the inverse of the channel coding rate, so $(lay_i) \times fec$ represents, in bytes, the increase of the i^{th} layer length when protected at a level γ . Unlike packet oriented

FEC scheme, where the 16 default RS codes are considered in the FEC rate allocation process, in this work we restrict the considered RS codes to those with $fec \leq 2$. In other words we only consider the first 10 default codes. This assumption make sense in layer oriented FEC rate allocation scheme because adding redundant data which in ratio is more than twice superior to the original layers may overload the networks and drastically increase the losses instead of reducing it.

Let γ be a layer protection level selected in the range $0 \leq \gamma \leq \gamma_{\max}^{lay}$, each protection level corresponds to a specific RS code selected between the 10 JPWL default RS codes ($\gamma=0$ means that the layer is not transmitted, $\gamma=1$ means transmission with protection level 1, higher values imply increasing channel code capacity with γ and $\gamma_{\max}^{lay} = 10$).

Let β_i be the number of data packet constituting the i^{th} quality layer of a JPEG 2000 codestream, $RD_{lay_i}^0$ and $RD_{lay_i}^\gamma$ be respectively the reduction of distortion associated to the correct decoding of layer i and the reduction of distortion associated to the correct decoding of layer i protected to level γ .

We rely on this codestream index file to derive $RD_{lay_i}^0$ and we associated the decoding error probability estimation process presented in (Yee et Weldon, 1995) in order to derive $RD_{lay_i}^\gamma$. Hence, the layer oriented FEC rate allocation problem is formalised by:

$$\text{Maximize } \sum_i^L \frac{RD_{lay_i, \gamma}}{(lay_i) \times fec_i} \quad (7)$$

$$\text{Subject to } \sum_i^L (lay_i) \times fec_i \leq B_{av} \quad (8)$$

3.2.b Optimization

We define $G_{lay_i}^\gamma$ as the gain in quality of the transmitted image obtained at the receiver side when layer i is decoded.

We derive $RD_{lay_i}^1$ and $RD_{lay_i}^\gamma$ the reduction of distortion obtained when layer i is transmitted respectively with protection level 1 and with protection level γ , we have:

$$\begin{aligned} RD_{lay_i}^1 &= (1 - P_{lay_i}^{i,1}) \times RD_{lay_i}^0 \\ RD_{lay_i}^\gamma &= (1 - P_{lay_i}^{i,\gamma}) \times RD_{lay_i}^0 \end{aligned} \quad (9)$$

Where $P_{lay_i}^1$ and $P_{lay_i}^\gamma$ are the decoding error probabilities obtained when layer i is protected respectively to level 1 and to level γ .

The resulting gain is:

$$G_{lay_i}^1 = \frac{RD_{lay_i,1}}{lay_i} = \frac{(1 - P_{lay_i}^1) \times RD_{lay_i}^0}{lay_i} \quad (10)$$

Similarly, any transmission between two consecutive protection levels ($\gamma-1$ and γ) yields an improvement in terms of reduction of distortion but has a budget cost equal to $(fec_\gamma - fec_{\gamma-1}) \times lay_i$, hence we have:

$$G_{lay_i}^1 = \frac{RD_{lay_i}^\gamma - RD_{lay_i}^{\gamma-1}}{(fec_\gamma - fec_{\gamma-1})lay_i}$$

$$G_{lay_i}^\gamma = \frac{(P_{lay_i}^{\gamma-1} - P_{lay_i}^\gamma)RD_{lay_i}^0}{(fec_\gamma - fec_{\gamma-1}) \times lay_i} \quad (11)$$

Applying the optimization process proposed in (Agueh et al, 2008), we derive the corresponding FEC rate for each layer. Since the Unequal Error protection is applied at layer level, the FEC rate is selected by decreasing order of layer importance. It is worth noting that all the packets belonging to the same layer are protected at the same FEC rate.

3.2.c Contribution of the optimal layer oriented FEC rate allocation scheme

Even if the gain metrics presented in the previous section seem close to the ones used in (Agueh et al, 2008), they hold a fundamental difference because they rely on the contribution of each layer to the reduction of distortion instead of just taking into account the contribution of a specific packet. Actually, during the source coding process the incremental contribution from the set of image codeblocks are collected in quality layers. Due to the fact that the rate-distortion compromises derived during JPEG 2000 truncation process are the same for all the codeblocks, for any quality layer index i the contributions of quality layer 1 through quality layer i constitute a rate-distortion optimal representation of the entire image. Hence, at layer level the reduction of distortion values are strictly decreasing. In contrast, the selection of a specific JPEG 2000 packet does not guarantee that the contributions of packet 1 to the selected index packet are monolithically decreasing. In this case, as confirmed by Descampe et al in (Descampe et al, 2006), some additional restrictions have to be added to the considered convex-hull in order to ensure rate-distortion and cost-distortion optimality. This justifies the necessity to have a merging step in the packet oriented FEC scheme (Agueh et al, 2008). Actually, it ensures that the convex-hull is always convex. In the layer oriented FEC this step is skipped because the reduction of distortion curve is already monolithically decreasing, significantly reducing the complexity and thus the time-consumption of the FEC rate allocation algorithm. Moreover, in the optimal layer oriented FEC scheme we only consider the first 10 RS codes instead of considering all the 16 default RS codes defined by JPWL standard as it is the case in (Agueh et al, 2008). This also considerably reduces the FEC scheme time consumption as it leads to less gains values computation which makes the proposed optimal layer FEC rate allocation scheme a good candidate for real time images/video streaming applications.

In addition, the number of layers available in the codestreams is another criterion which contributes to the reduction of the time consumption of our proposed FEC scheme. Actually, a JPEG 2000 image extracted from a Motion JPEG 2000 video sequence is defined by (L, R, C) where L is the number of quality layers of the considered image, R is its resolution level corresponding to the decomposition levels of the Discrete Wavelet Transform and C is the number of components. Assuming that the considered JPEG 2000 image is not spatially divided and thus is described by a unique tile, the number of data packets available in the considered JPEG 2000 codestreams is given by $S = L \times R \times C$. In this context, the complexity of packet oriented FEC schemes is based on the S data packets while the complexity of the optimal layer based FEC is based on the L layers available in the

codestreams. In scalable JPEG 2000 images, since the number of layers is significantly lower in comparison to the number of data packets, the time consumption of our proposed layer oriented FEC scheme is significantly low in comparison to packet oriented scheme.

3.2.d Algorithm

For each JPEG 2000 image

- Model the channel with a Gilbert model and for each possible protection level γ ($0 \leq \gamma \leq 10$), evaluate the probability of incorrect word decoding $P_{lay_i}^\gamma$

- For $i = 1$ to $i = L$ (Number of JPEG 2000 layers)

For $\gamma = 1$ to $\gamma = 10$

Estimate $RD_{lay_i}^\gamma = (1 - P_{lay_i}^{\gamma}) \times RD_{lay_i}^0$

$$G_{lay_i}^1 = \frac{RD_{lay_i}^\gamma - RD_{lay_i}^{\gamma-1}}{(fec_\gamma - fec_{\gamma-1})lay_i}$$

End For

End For

- Ordering gain values in decreasing order of importance
- Selecting each gain value, corresponding to a specific protection level, up to meeting the rate constraint
- Optimally protect JPEG 2000 layers with the corresponding RS codes

End For

3.2.e Performance of layer based FEC scheme in terms of time consumption

In Figure 6 the run time of the proposed layer based FEC rate allocation scheme is plotted versus the number of data packets available in the JPEG 2000 codestreams. This curve is compared to the one achieved using the optimal packet oriented FEC rate allocation scheme (Agueh et al, 2008). These results are achieved using an Intel core Duo CPU 2.9 Ghz Workstation.

As packet-oriented and layer oriented schemes are linked by the number of layers available in each image, we vary this parameter in order to derive some comparable results. In the considered scenario, the number of available resolution and component of JPEG 2000 frames are fixed (resolution = 10 and component = 1) because these parameters do not impact the time-performance of layer oriented FEC rate allocation schemes. In Figure 6 each packet (i) corresponds to a specific JPEG 2000 frame (with a specific quality layer).

In this scenario, the available bandwidth in the system is set to 18 Mbits/s ($B_{av} = 18 \text{ Mbits} / \text{s}$). It is worth noting that in practice few existing JPEG 2000 codecs allow high quality scalability and to our knowledge, none of them can handle more than 50 quality layers. Hence, the considered scenario allows generalization to future high quality layer scalable FEC rate allocation systems.

In Figure 6 we notice that both layer and packet oriented scheme have a run time linearly increasing with the number of packets available in the codestreams. However, the optimal layer based FEC scheme is significantly less time consuming than the packet based FEC scheme. For codestreams containing less than 1000 packets (quality layers ≤ 10) the packet oriented FEC scheme is 3 times more time consuming than our optimal layer based FEC scheme. For JPEG 2000 codestreams, whose number of packets is between 1000 and 5000

(quality layers between 10 and 50) the packet oriented scheme is up to 5 times the run time of the layer based FEC scheme. Since existing JPEG 2000 codecs handle less than 50 quality layers, our proposed optimal layer based scheme is a good candidate for real-time JPEG 2000 codestreams over wireless channel as it yields low time consumption.

The proposed optimal layer based scheme, due to its low time consumption, could be viewed as a good candidate for future high quality layer scalable wireless JPEG 2000 based images and video streaming applications.

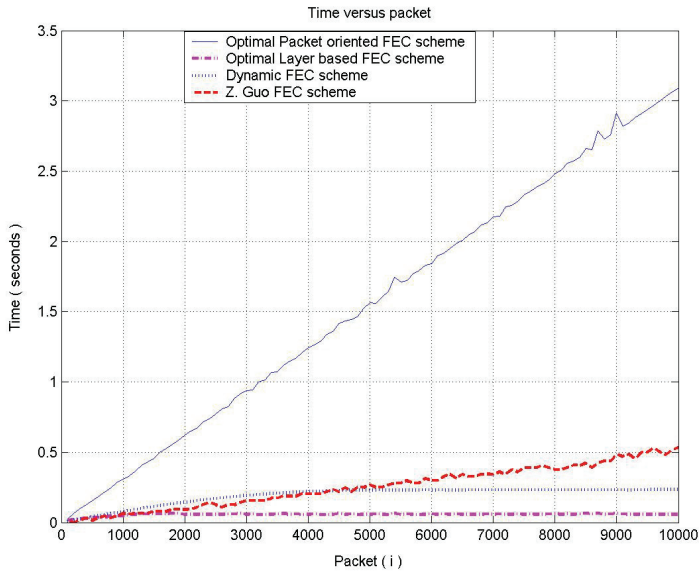


Fig. 6. Time versus packets: Fixed image resolution ($R=10$) -Varying quality layers (0 to 100) - One component ($C=1$)

Although the layer based scheme achieves good performances in terms of time consumption in comparison to packet oriented FEC rate allocation schemes, the last ones present better performance in terms of visualization quality especially for highly noisy channels. In the following section we demonstrate the effectiveness of the optimal layer based FEC scheme thanks to a client/server application of Motion JPEG 2000 video streaming over real ad-hoc network traces.

3.2.f Packet-oriented and Layer-oriented FEC rate allocation for Motion JPEG 2000 video streaming over real ad-hoc network traces

In this section we present the results achieved while streaming Motion JPEG 2000 based video over real ad-hoc network channel traces (Loss patterns acquired during the WCAM Ancey 2004 measurement campaigns IST-2003-507204 WCAM, Wireless Cameras and Audio-Visual Seamless Networking, 2004) and we demonstrate that the proposed optimal layer based scheme outperforms existing layer oriented FEC schemes even if for highly noisy channel it is less efficient than packet oriented FEC scheme. The comparison is handled both in terms of Structural Similarity (SSIM) (Wang et al, 2004) and in terms of successful decoding rate. We derive the Mean SSIM metric of the Motion JPEG 2000 video

sequence by averaging the SSIM metrics of the JPEG 2000 images contained on the considered video sequence. It is worth noting that each SSIM measure derived is associated to a successful decoding rate metric which corresponds to decoder crash avoidance on the basis of 1000 transmission trials.

The considered wireless channel traces are available in (Loss Patterns, 2004) and the video sequence used is *speedway.mj2* (Speedway, 2005) containing 200 JPEG 2000 frames generated with an overall compression ratio of 20 for the base layer, 10 for the second layer and 5 for the third layer. Figure 7 presents the successful decoding rate of the motion JPEG 2000 video sequence *speedway.mj2* (Speedway, 2005) transmission over real ad-hoc network channel traces (Loss Patterns, 2004). We observe that for highly noisy channels ($C/N \leq 15$ dB), the proposed optimal layer outperforms other layer based FEC schemes but is less efficient than the packet oriented scheme. For noisy channel (15 dB $\leq C/N < 18$ dB), we notice that all layer based UEP schemes exhibit similar performances in terms of successful decoding rate. For low noisy channel ($C/N \geq 18$ dB) all the FEC schemes yield the same improvement in terms of successful decoding rate.

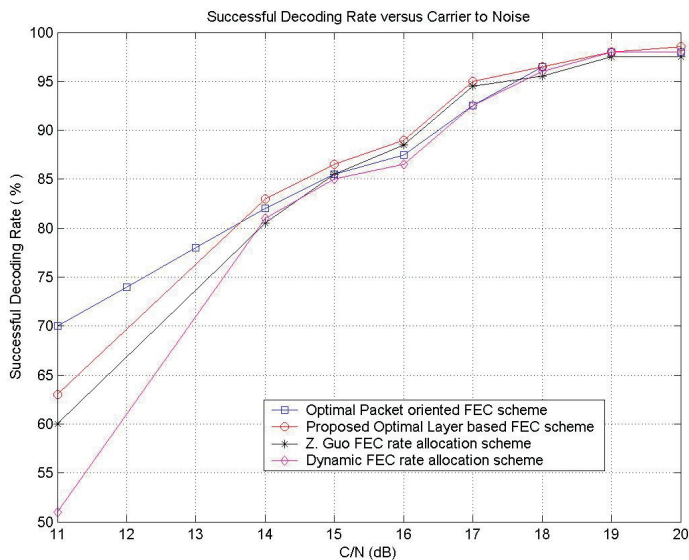


Fig. 7. Successful decoding rate versus Carrier to Noise Ratio

In Figure 8 we show that our proposed optimal layer based FEC rate allocation scheme still outperforms other layer based schemes in terms of Mean SSIM. This is due to the fact that the base layer which is the most important part of the codestreams is highly protected in our proposed scheme, in comparison to other layer based schemes, guaranteeing this way a good quality for the visualization.

It is worth noting that, for highly noisy channels, our optimal layer oriented FEC scheme is less efficient than optimal packet oriented FEC scheme presented in (Agueh et al, 2008). However the last one is unpractical for real time streaming applications when the number of packets in the codestreams is high. In contrast our proposed layer oriented efficiently overcomes this limitation. In this context, instead of being used to replace packet oriented FEC rate allocation schemes, our proposed optimal layer based FEC scheme should be used to complete it.

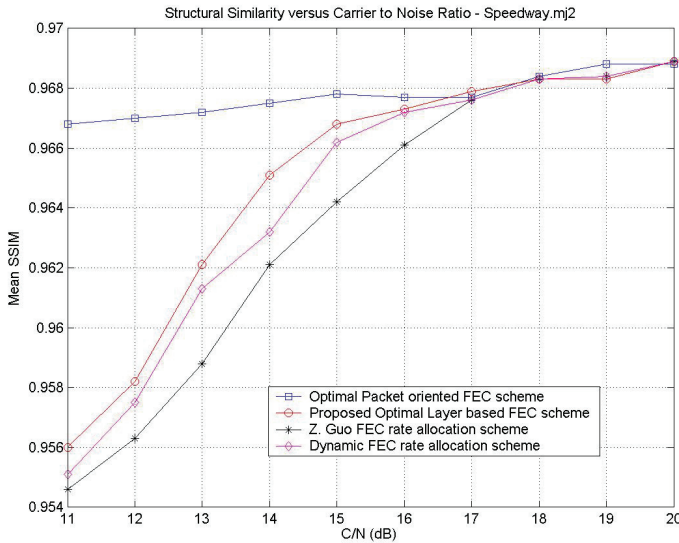


Fig. 8. Mean Structural Similarity versus Carrier to Noise Ratio

4. Discussion and prospective issues

4.1 On JPEG 2000 codestreams interleaving

In this section, we discuss and evaluate the impact of data interleaving in the effectiveness of the FEC rate allocation scheme. Thanks to the interleaving matrix presented in Figure 9, protected JPEG 2000 data are decorrelated before being sent through the wireless channel. Hence, the impact of consecutive channel errors sequences on the transmitted codestreams is reduced. In Figure 9 the protected JPEG 2000 codestream is divided into Px packets of length N . Then, the interleaving process consists in storing M consecutive packets into a $M \times N$ matrix and to read the columns of this matrix so that two initially consecutive symbols are separated by a distance of $I = M$ (symbols). We refer to I as the interleaving degree. The considered channel is a real mobile ad-hoc network channel experiencing $PER = 3.88 \times 10^{-2}$ and the interleaving degrees are 1, 2, 4, 8, 16, 32, 64 and 128. Table 1 shows the PSNR evolution as function of interleaving degree I . The considered image is *speedway_0.j2k* protected with the optimal packet-oriented JPWL compliant scheme.

The interest of interleaving is shown in table 1 in the sense that the PSNR and the successful decoding rate increase with the interleaving degree I . The results in table 1 are valid for a Gilbert channel with a specific error correlation factor and are no longer the same when this factor changes. For the considered channel, we observe that for $I \leq 8$, interleaving has no noticeable impact because the interleaving degree I is smaller than the average error burst length. In fact, we show in (Agueh et al, 2008) that the upper bound of the mean error burst length is $L_B^{max} = 10$ bytes. Hence, in order to be efficient, the interleaving degree should be higher than 10 bytes. When I is increased to 16 or more, we notice an improvement of both the PSNR and the successful decoding rate. However, we observe that higher values of I

(128) yield only slight improvement in terms of PSNR while consuming considerable memory resources leading to the conclusion that reasonable interleaving degree (typically $I = 16$ or $I = 32$) is a good compromise.

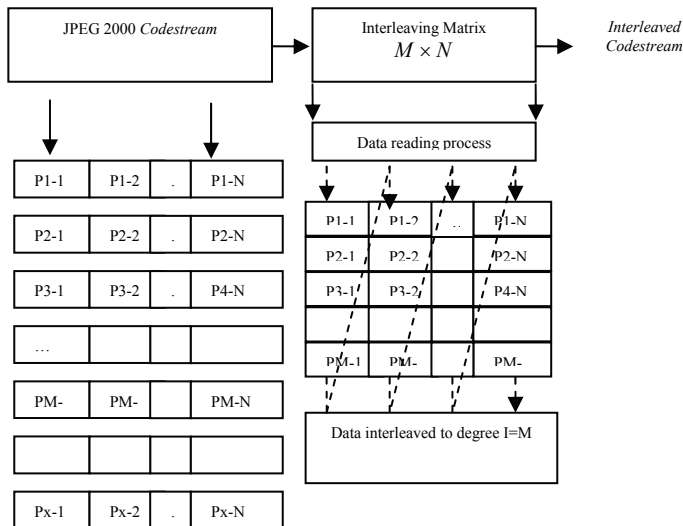


Fig. 9. Interleaving process

Interleaving degree I	PSNR (dB)	Successful Decoding Rate
$I=1$	24.1	77.5
$I=2$	24.6	89.8
$I=4$	25.2	92.1
$I=8$	31.8	93.4
$I=16$	38.7	94.5
$I=32$	44.33	94.7
$I=64$	44.38	94.9
$I=128$	44.37	94.8

Table 1. Interleaving degree and associated image PSNR

Even if the empirical selection of interleaving degree is commonly accepted, it is worth noting that optimal dynamic selection of interleaving degree is still an interesting and open issue. Moreover, proposing new interleaving schemes which are able to take into account the specificity of JPEG 2000 codestreams is also an open issue. Fabrizio & Baruffa address this issue in (Fabrizio & Baruffa, 2005) by proposing a backward-compatible interleaving technique for robust JPEG 2000 wireless transmission. To the best of our knowledge, the virtual interleaving scheme proposed by Fabrizio & Baruffa is the only JPEG 2000 based backward compatible interleaving scheme available in the literature. Hence, original and new interleaving scheme are still needed to improve the robustness of JPEG 2000 codestreams against transmission errors in wireless networks.

4.2 Scalable JPEG 2000 transmission

Many problems are still to be addressed in the framework of JPEG 2000 codestreams transmission over wireless networks. Image scalability based on dynamic available bandwidth estimation is one of those problems. In the literature, proposed image scalable systems have been implemented using a fixed available bandwidth in their considered scenarios (Li and Chang, 2009), (Devaux et al, 2007). This assumption is no longer true in wireless systems because they rely on radio waves whose characteristics depend on the transmission environment. Moreover, few of the proposed systems addressed simultaneously the bandwidth estimation problem and the issue of smoothness for JPEG 2000 codestreams scalability. In (Mairal & Agueh, 2010), we address both issues by proposing a scalable and non aggressive wireless JPEG 2000 image and video transmission algorithm based on a dynamic bandwidth estimation tool.

The main limitation of the scalable system proposed in (Mairal & Agueh, 2010) is that it handles only one wireless client (see section 4.2.a). However, this limitation could be overcome by generalizing the proposed algorithm to multiple wireless clients' scenario. We propose in section 4.2.b, a framework for this generalization which opens the way for efficient wireless JPEG 2000 codestreams transmission in Next Generation Networks which are characterized by the cohabitation of multiples wireless devices having different standards requirements and different capacities.

4.2.a Scalable JPEG 2000 transmission for single wireless client

In this section we present an adaptive bandwidth estimation tool and propose an additional scalability tool at the encoder, which dynamically and efficiently selects the best resolution and layer for each JPEG 2000 frame before transmission through the wireless channel. Hence, according to the estimated bandwidth, refinement layers could be added or removed from JPEG 2000 codestreams. When required, scale changes are gradual and smooth in order to guarantee a comfort in terms of visualization. We present in the following the processes, which are implemented at the encoder.

Algorithm:

Once connected, the server starts the WBest process in order to obtain the initial available bandwidth. WBest is the available bandwidth estimation tool implemented in the system. A detailed description of WBest is provided in (Li and Chang, 2009). At this step the goal is to send images and video with maximum detail (highest resolution and all refinement layers) matching with the estimated bandwidth. The original resolution and number of layers of the considered video is found using an indexer like the one available in OpenJpeg (www.openjpeg.org). In (Mairal & Agueh, 2010), the default number of resolutions is 6, the length and the width of the image must be a power of 2 (here 352x288), the number of layers is 3. Let l be a layer of a JPEG 2000 image and SE_{rate}^l is corresponding source encoding rate. Let $fec_{rate}^l = \frac{n}{k}$ be the inverse of the Reed-Solomon code RS(n,k) selected by the FEC rate allocation scheme to protect layer l against transmission errors. Let $frame_length$ be the amount of data needed to transmit layer l protected. We have:

$$frame_length = H \times W \times SE_{rate}^l \times fec_{rate}^l \quad (12)$$

The proposed scheme is able to adapt to channel conditions thanks to the bandwidth estimation tool. Hence, when the channel experienced good conditions, our heuristic selects

the highest resolution with the highest quality (all the refinement layers are transmitted). If the channel experienced harsh conditions, image layers and resolution are decreased up to defined thresholds. We empirically set thresholds ($l_{\max} / 2$) and ($resol_{\min}^{desired}$) as respectively the minimum layer downscaling allowed and the minimal resolution, which guaranties comfort in terms of image visualization. Contrarily, l_{\min} and $resol_{\min}$, respectively the base layer and the minimal resolution possible do not guaranty a visual comfort. Hence, when the channel experienced bad conditions, image layers are incrementally reduced while maintaining original resolution of the JPEG 2000 frame to highest level. However, is the corresponding frame length do not match the available bandwidth, image resolution downscaling is processed. It is worth noting that our fixed thresholds are valid for our scenarios and may change in different environments.

Algorithm

- 1-Wait for client connection request
- 2-Estimate available bandwidth (B_{av}) using WBest tool every 10 frames
- 3-Derive original images/video maximum number of layers (l_{\max}) and maximum resolution ($resol_{\max}$) from the indexer
- For each JPEG 2000 image:
 - 4-Initialize parameters current layer ($cur_lay = l_{\max}$) and current resolution ($cur_resol = resol_{\max}$)
 - 5- Calculate the needed bandwidth B_{needed}
 - 6- If ($B_{av} \geq B_{needed}$) Send Image
 - Else
 - {
 - Step 1: Guaranty of comfort during visualization
 - While ($cur_lay > l_{\max} / 2$ and $cur_resol > resol_{\min}^{desired}$)
 - { Estimate $B_{needed} = \sum_{i=1}^{cur_lay} cur_resol \times SE_{rate}^i \times fec_{rate}^i$
 - Increase layer to remove ($lay_to_rem = lay_to_rem + 1$) and
 - Fixe resolution or fixe layer to remove and decrease resolution ($cur_resol = cur_resol - 1$) until reaching ($B_{av} \geq B_{needed}$)
 - }
 - }
 - Step 2: Without guaranty of comfort during visualization
 - While ($cur_lay > 0$ and $cur_resol > resol_{\min}$)
 - { Estimate $B_{needed} = \sum_{i=1}^{cur_lay} cur_resol \times SE_{rate}^i \times fec_{rate}^i$
 - Increase layer to remove ($lay_to_rem = lay_to_rem + 1$) and
 - Fixe resolution or fixe layer to remove and decrease resolution ($cur_resol = cur_resol - 1$) until reaching ($B_{av} \geq B_{needed}$)
 - }
 - }
 - 7- Smooth scale changes in order to avoid sudden and temporal image variation (average among 5 previous frames parameters)

Once the resolution and the number of layers are chosen, the server sends the video streaming to the client.

The available bandwidth estimation tool is launched every 10 frames but this frequency could be changed according to the application requirements.

In our work, we evaluate the extra time yielded by different refreshment frequencies while transmitting *speedway.mj2* video. We have:

$$\Delta t(\text{refr}) = 75e^{-0.115\text{refr}} \text{ (seconds)}$$

Where Δt and refr are respectively the extra time and the number of transmitted frames between two consecutive bandwidth estimations. We derive from this study that a refreshment of 10 frames increases less than 20% the total transmission time while guaranteeing sufficiently accurate channel tracking.

An interesting extension to this work could be to optimally adapt the frequency of the bandwidth estimation tool to the channel conditions.

The efficiency of the proposed heuristic is demonstrated using a wireless client/server video streaming application. In the following section, we present the results derived from different video streaming scenarios.

The video streaming scenarios considered in this work are derived from wireless transmission trials used in the literature for bandwidth estimation purpose. *WBest* is the available bandwidth estimation tool implemented in our system.

The video sequence is *speedway.mj2*, which is a 352x288 motion JPEG2000 sequence constituted of 200 JPEG 2000 frames with six resolutions and three layers each.

Scenario

In the considered scenario, the wireless channel considered is derived from BART tool (Johnsson & Björkman, 2008), which estimates the available bandwidth in an end-to-end path where the bottleneck is a wireless hop. It uses the Probing Packet Pair Trains Dispersion Technique and improves the system using Kalman filters to measure and track the changes.

In this scenario, we focus on the fast varying part of the estimated bandwidth. Moreover, we divide the bandwidth estimated by BART tool 0. by a parameter $\delta=2$ in order to show that our scheme is efficient even when the wireless channel experienced harsh conditions.

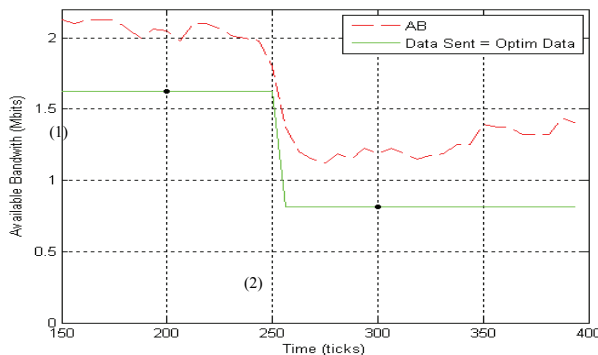


Fig. 10. Available bandwidth versus time - BART Tool scenario

Scenario	Image Length	Image Width	Layer of transmitted Images
(1)	352	288	3
(2)	352	288	2

Table 2. Scalability parameters - BART Tool scenario

In Figure 10, point (1) indicates that the estimated bandwidth is higher than the needed bandwidth, hence initial JPEG 2000 frames are transmitted. Point (2) shows that the estimated bandwidth is decreased and becomes insufficient to send original images. Hence, the algorithm maintains the resolution at the highest level (initial value) but one layer is removed from original frames as shown in Table 2.

In this section, the performance of scalable JPEG 2000 transmission over wireless networks is outlined. In the following section we address the problem of JPEG 2000 codestreams scalability in a context of multiple wireless clients networking.

4.2.b Scalable JPEG 2000 transmission for multiple wireless clients

In order to address the issue of simultaneous service provisioning for multiple wireless receivers, a multithread server is implemented at the encoder.

After detecting the value of Max_Client the number of wireless receivers in a considered cell, the server waits for client's connection requests. Once connected and identified, a receiver sends the maximal resolution of its viewer to the server so that the adequate size and image quality layers could be selected. Then, the available bandwidth estimation tool is launched to obtain an initial estimation of the downlink capacity. After the available bandwidth estimation, the algorithm presented in (Mairal & Agueh, 2010) selects the suitable size and number of layers of the images to be sent.

In the following we present the processes, which are implemented at the encoder.

Algorithm:

1. While (number of connected clients ($connected_clients$) \leq Max_Client)
 - {
 - 2. Server waits for a client to connect to his socket
 - 3. Server starts a new thread to serve the new wireless client
 - Increment the number of connected clients ($connected_clients$);
 - }
 - 4. Receive wireless client equipment resolution ($max_device_resolution$)
 - 5. While (transmitting JPEG 2000 images/video)
 - {
 - Every 10 frames
 - {
 - Launch WBest and estimate available Bandwidth (B_{av})
 - Select the suitable scalability parameters
 - }
 - Send frames with selected parameters
 - Increment the number of transmitted frames ($Sent_frames$);
 - }
 - }

The selected size is never bigger than the wireless device's resolution. The clients connected to the server receive the highest images/video quality achievable for the estimated available bandwidth. The available bandwidth estimation tool is launched every 10 frames but this frequency could be changed according to the application requirements. As recommended in (Mairal & Agueh, 2010), we fix this frequency to 10 frames. The result is a smooth and robust video sequence for each receiver. It is worth noting that the original JPEG 2000 codestreams is copied and stored at the encoder so that new JPEG 2000 codestreams are generated for each client according to the parameters selected by the proposed scalability algorithm. An interesting extension to this work is to implement a real time system which is able to handle real time multiple JPEG 2000 source coding.

Scenario

The video streaming scenarios considered in this work are derived from wireless transmission trials used in the literature for bandwidth estimation purpose. WBest is the available bandwidth estimation tool implemented in our system.

The video sequence is *speedway.mj2* (Speedway, 2005) which is a 352x288 motion JPEG2000 sequence constituted of 200 JPEG 2000 frames with six resolutions and three layers each.

In the considered scenario (figure 11), an IEEE 802.11 based wireless network with two receivers and one sender connected to an access point is used to demonstrate the effectiveness of the proposed system.

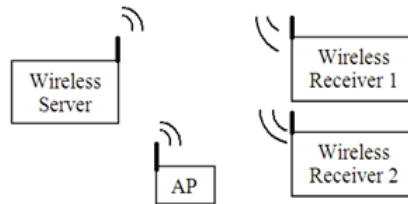


Fig. 11. Single server / multiple wireless clients' transmission scenario

In Table 3, we present the results of the considered scenario where *speedway.mj2* is transmitted to multiple wireless clients. The estimated available bandwidths along with the characteristics of the selected stream for each client are presented. In the first case, the client receives a video composed of JPEG 2000 frames with all layers and which size is 176x144.

In this scenario, the size of transmitted video sequence is decreased in comparison to the original JPEG 2000 codestreams because the estimated available bandwidth is not higher enough to allow full original video transmission. The proposed algorithm helps preserving the details by guarantying a comfort of visualization. As we recommend in (Mairal & Agueh, 2010), when more than half of the original layers have to be removed to fit bandwidth requirements, images quality layers is maintained while their size is decreased. Hence, for the first receiver, instead of transmitting JPEG 2000 images with 352x176 sizes with only the base layer, the server transmits images with all the layers after decreasing the resolution. It is worth noting that in the considered scenario, maintaining the same size as user equipment settings is not mandatory.

For the second receiver (see Table 3), the video transmitted is a sequence of JPEG 2000 frames with the all the layers, but with a decreased resolution of 88x72.

The third column of Table 3, indicates the channel occupation during the transmission. We can observe that the available bandwidth is never exceeded.

Frames	Parameters	Receiver 1	Receiver 2	Total
0 to 40	Estimated Available BW	744 kbits	704 kbits	1448 kbits
	Amount of data sent	405504 bits	101376 bits	506880 bits
	Number of layers	3	3	--
	Resolution	176x144	88x72	--
40 to 199	Estimated Available BW	704 kbits	704 kbits	1408 kbits
	Amount of data sent	405504 bits	101376 bits	506880 bits
	Number of layers	3	3	--
	Resolution	176x144	88x72	--

Table 3. Available bandwidth and data sent for both receivers

Thanks to wireless clients/server applications, we demonstrate the efficiency of the proposed scalable system. Even if the performance of the proposed system still to be evaluated in harsh wireless environments, the proposed scalable system could be viewed as a valid step toward guaranteeing Quality of Service (QoS) for JPEG 2000 based multimedia transmission in heterogeneous Next Generation Wireless networks.

5. References

- Taubman, D.S. & Marcellin, M.W. (2001). JPEG 2000 Image Compression Fundamentals, Standards and Practice, In: Kluwer Academic Publishers, The Netherlands 2001
- JPWL, (2005) JPEG 2000 part 11 Final Draft International Standard, ISO/IEC JTC 1/SC 29/WG 1 N3797
- Dufaux, F. & Nicholson, D. (2004). JPWL: JPEG 2000 for Wireless Applications, *Proceeding of SPIE -- Volume 5558 - Applications of Digital Image Processing XXVII*, Andrew G. Tescher, Editor, pp. 309-318, November 2004
- Nicholson, D., Lamy-Bergot, C., Naturel, & Poulliat, C. (2003). JPEG 2000 backward compatible error protection with Reed-Solomon codes. *IEEE Transactions on Consumer Electronics*, vol. 49, n. 4, pp.855-860, Nov. 2003
- Agueh, M., Devaux, F.O. & Diouris, J.F. (2007,a). A Wireless Motion JPEG 2000 video streaming scheme with a priori channel coding. *Proceeding of 13th European Wireless 2007 (EW-2007)*, April 2007, Paris (France)

- Guo, Z., Nishikawa, Y., Omaki, R. Y., Onoye, T. & Shirakawa, I. (2006). A Low-Complexity FEC Assignment Scheme for Motion JPEG 2000 over Wireless Network. *IEEE Transactions on Consumer Electronics*, Vol. 52, Issue 1, Feb. 2006 Page(s): 81 – 86
- Agueh, M., Diouris, J.F., Diop, M., & Devaux, F.O. (2007,b). Dynamic channel coding for efficient Motion JPEG 2000 streaming over MANET. *Proceeding of Mobimedia2007*, August 2007, Nafpaktos, Greece
- Agueh, M., Diouris, J.F., Diop, M., Devaux, F.O., De Vleeschouwer, C. & Macq, B. (2008). Optimal JPWL Forward Error Correction rate allocation for robust JPEG 2000 images and video streaming over Mobile Ad-hoc Networks. *EURASIP Journal on Advances in Signal Proc.*, Spec. Issue wireless video, Vol. 2008, Article ID 192984, doi: 10.1155/2008/192984
- Agueh, M., Ataman, S. & Henoc, S. (2009, a). A low time-consuming smart FEC rate allocation scheme for robust wireless JPEG 2000 images and video transmission. *Proceeding of Chinacom2009*, 2009, Xi'an(China)
- Agueh, M. & Henoc, S. (2009, b). Optimal Layer-Based Unequal Error Protection for Robust JPEG 2000 Images and Video Transmission over Wireless Channels. *Proceeding of MMEDIA2009*, pp.104– 109, 2009 First International Conference on Advances in Multimedia, 2009
- Schulzrinne, H. , Casner, S. , Frederick, R. & Jacobson, V. (2003). RTP: A Transport Protocol for Real-Time Applications. STD 64, RFC 3550, July 2003.
- Thomos, N., Boulgouris, N. V. & Strintzis, M. G. (2004). Wireless transmission of images using JPEG 2000. *Proceeding of ICIP'04*, Singapore, October 2004.
- Speedway video sequences have been generated by UCL.
Available: <http://euterpe.tele.ucl.ac.be/WCAM/public/Speedway%20Sequence/>
- Loss patterns acquired during the WCAM Annecy 2004 measurement campaigns IST-2003-507204 WCAM, Wireless Cameras and Audio-Visual Seamless Networking.
project website: <http://www.ist-wcam.org>
- Yee, J. R., & Weldon E. J. (1995). Evaluation of the performance of error-correcting codes on a Gilbert channel. *IEEE Transactions on Communications*. vol. 43, n. 8, pp. 2316-2323, 1995
- Descampe, A., De Vleeschouwer, C., Iregui, C., Macq, B. & Marques, F. (2006). Pre-fetching and caching strategies for remote and interactive browsing of JPEG 2000 images. *IEEE Transactions on Image Processing*, vol 16, n° 5, pp. 1339-1354, 2006
- Wang, Z., Bovik, A. C. , Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions. on Image Processing.*, vol. 13,no. 4, pp 600-612, Apr. 2004
- Fabrizio, F. & Giuseppe, B. (2005). Backward-compatible interleaving technique for robust JPEG2000 wireless transmission Visual content processing and representation. *Proceeding of 9th international workshop, VLBV 2005*, Sardinia, Italy, September 15-16, 2005-VLBV 2005 N°9, Sardinia, ITALIE (2005)
- Li, M. & Chang, C. (2009). A two-way available bandwidth estimation scheme for multimedia streaming networks adopting scalable video coding. *Proceeding of IEEE Sarnoff Symposium*, p1-p11, Princeton, USA, 2009.

- Devaux, F., Meessen, J., Parisot, C., Delaigle, J., Macq, B. & De Vleeschouwer, C. (2007). A flexible video transmission system based on JPEG 2000 conditional replenishment with multiple reference. Proceeding of IEEE ICASSP 2007, Honolulu 2007, USA
- Mairal, C. & Agueh, M. (2010). Smooth and Scalable Wireless JPEG 2000 images and video streaming with dynamic Bandwidth Estimation. *Proceeding Of the Second International conference on advances in Multimedia*, June 2010, Athens, Greece
- Johnsson, A. & Björkman, M. (2008). On measuring available bandwidth in wireless networks. Proceeding of 33rd IEEE Conference on Local computer networks, 2008. Montreal, Canada

Downlink Capacity of Distributed Antenna Systems in a Multi-Cell Environment

Wei Feng, Yunzhou Li, Shidong Zhou and Jing Wang
State Key Laboratory on Microwave and Digital Communications
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084,
P. R. China

1. Introduction

The future wireless networks will provide data services at a high bit rate for a large number of users. Distributed antenna system (DAS), as a promising technique, has attracted worldwide research interests (Feng et al., 2010; 2009; Saleh et al., 1987; Ni & Li, 2004; Xiao et al., 2003; Zhuang et al., 2003; Hasegawa et al., 2003; Choi & Andrews, 2007; Roh & Paulraj, 2002b;a). In DAS, the antenna elements are geographically separated from each other in the coverage area, and the optical fibers are employed to transfer information and signaling between the distributed antennas and the central processor where all signals are jointly processed. Recent studies have identified the advantages of DAS in terms of increased system capacity (Ni & Li, 2004; Xiao et al., 2003; Zhuang et al., 2003; Hasegawa et al., 2003; Choi & Andrews, 2007) and macro diversity (Roh & Paulraj, 2002b;a) as well as coverage improvement (Saleh et al., 1987).

Since the demand for high bit rate data service will be dominant in the downlink, many studies on DAS have focused on analyzing the system performance in the downlink. In a single cell environment, the downlink capacity of a DAS was investigated in virtue of traditional MIMO theory (Ni & Li, 2004; Xiao et al., 2003; Zhuang et al., 2003). However, these studies did not consider per distributed antenna power constraint, which is a more practical assumption than total transmit power constraint. Moreover, the advantage of a DAS should be characterized in a multi-cell environment. (Hasegawa et al., 2003) addressed the downlink performance of a code division multiple access (CDMA) DAS in a multi-cell environment using computer simulations, but it did not provide theoretical analysis. A recent work (Choi & Andrews, 2007) investigated the downlink capacity of a DAS with per distributed antenna power constraint in a multi-cell environment and derived an analytical expression. However, it was only applicable for single-antenna mobile terminals.

In this chapter, without loss of generality, the DAS with random antenna layout (Zhuang et al., 2003) is investigated. We focus on characterizing the downlink capacity with the generalized assumptions: (a1) per distributed antenna power constraint, (a2) generalized mobile terminals equipped with multiple antennas, (a3) a multi-cell environment. Based on system scale-up, we derive a quite good approximation of the ergodic downlink capacity by adopting random matrix theory. We also propose an iterative method to calculate the

unknown parameter in the approximation. The approximation is illustrated to be quite accurate and the iterative method is verified to be quite efficient by Monte Carlo simulations.

The rest of this chapter is organized as follows. The system model is described in the next section. Derivations of the ergodic downlink capacity are derived in Section 3. Simulation results are shown in Section 4. Finally, conclusions are given in Section 5.

Notations: Lower case and upper case boldface symbols denote vectors and matrices, respectively. $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and the transpose conjugate, respectively. $\mathbb{C}^{M \times N}$ represents the complex matrix space composed of all $M \times N$ matrices and \mathcal{CN} denotes a complex Gaussian distribution. $\mathbf{E}(\cdot)$ and $\text{Var}(\cdot)$ represent the expectation operator and variance operator, respectively. \mathbf{I}_n is an identity matrix with the dimension equal to n . “ \otimes ” denotes the Kronecker product.

2. System model

The architecture of a DAS with random antenna layout in a multi-cell environment is illustrated in Fig. 1, where a cell is covered by N uniformly-distributed antennas (DAs), and each cell is loaded with a single randomly-deployed mobile terminal¹ (MT), which is equipped with M antenna elements (AEs). The optical fibers are employed to transfer information and signaling between the central processor and the DAs.

We consider the 1-tier cellular structure (Choi & Andrews, 2007) with universal frequency reuse, where a given cell (indexed by $i = 0$) is surrounded by one continuous tier of six cells (indexed by $i = 1 \sim 6$) as shown in Fig. 1.

Basically, the downlink of the considered DAS is a $N \times M$ MIMO system with interference and noise. The received signal vector of the terminal in the 0th cell can be expressed as (Telatar, 1999)

$$\begin{aligned} \mathbf{y}^{(0)} &= (\text{signal}) + (\text{interference}) + (\text{noise}) \\ &= \mathbf{H}^{(0)} \mathbf{x}^{(0)} + \sum_{i=1}^6 \mathbf{H}^{(i)} \mathbf{x}^{(i)} + \mathbf{n}, \end{aligned} \quad (1)$$

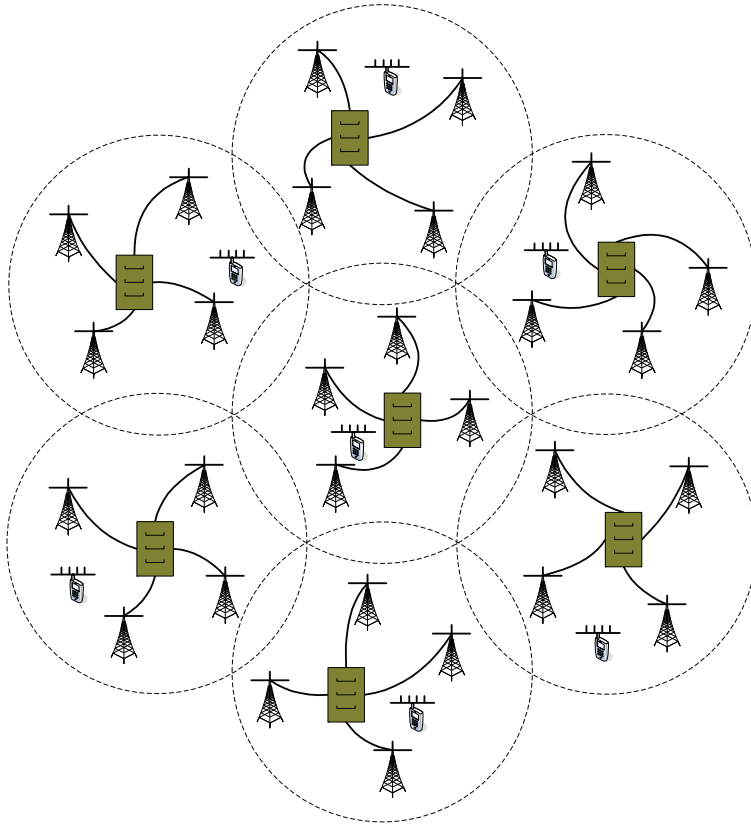
where $\mathbf{H}^{(i)} \in \mathbb{C}^{M \times N}$, $i = 0, 1, \dots, 6$, denotes the channel matrix between the DAs in the i th cell and the MT in the 0th cell, $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_N^{(i)}] \in \mathbb{C}^{N \times 1}$, $i = 0, 1, \dots, 6$, is the transmitted signal vector of the DAs in the i th cell, $\mathbf{n} \in \mathbb{C}^{M \times 1}$ denotes the white noise vector with distribution $\mathcal{CN}(0, \sigma_n^2 \mathbf{I}_M)$. The per distributed antenna power constraint is considered, we have

$$\mathbf{E} \left[|x_n^{(i)}|^2 \right] \leq P_n^{(i)}, n = 1 \sim N, i = 0 \sim 6, \quad (2)$$

where $P_n^{(i)}$ denotes the power constraint of the n th DA in the i th cell.

The composite fading channel matrix $\mathbf{H}^{(i)}$, $i = 0, 1, \dots, 6$, encompasses not only small-scale fading but also large-scale fading (Roh & Paulraj, 2002b;a), which is modeled as

¹ The system corresponds to the set of MTs using a particular orthogonal dimension, e.g., a time slot for time division multiple access.








 Distributed Antenna
  Mobile Terminal
 Antenna Element
  Antenna Elements
  Central Processor

Fig. 1. Illustration of a DAS in a multi-cell environment.

$$\begin{aligned}
 \mathbf{H}^{(i)} &= \mathbf{H}_w^{(i)} \mathbf{L}^{(i)} \\
 &= \begin{bmatrix} h_{11}^{(i)} & \dots & h_{1N}^{(i)} \\ \vdots & \ddots & \vdots \\ h_{M1}^{(i)} & \dots & h_{MN}^{(i)} \end{bmatrix} \begin{bmatrix} l_1^{(i)} & & \\ & \ddots & \\ & & l_N^{(i)} \end{bmatrix} \\
 &= \begin{bmatrix} h_{11}^{(i)} l_1^{(i)} & \dots & h_{1N}^{(i)} l_N^{(i)} \\ \vdots & \ddots & \vdots \\ h_{M1}^{(i)} l_1^{(i)} & \dots & h_{MN}^{(i)} l_N^{(i)} \end{bmatrix},
 \end{aligned} \tag{3}$$

where $\mathbf{H}_w^{(i)}$ and $\mathbf{L}^{(i)}$ reflect the small-scale channel fading and the large-scale channel fading between the DAs in the i th cell and the MT in the 0th cell, respectively.

$\{h_{mn}^{(i)} \mid m=1,2,\dots,M; n=1,2,\dots,N; i=0,1,\dots,6\}$ are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian variables with zero mean and unit variance, and $\{l_n^{(i)} \mid n=1,2,\dots,N; i=0,1,\dots,6\}$ can be modeled as

$$l_n^{(i)} = \sqrt{\left[D_n^{(i)}\right]^{-\gamma} S_n^{(i)}}, n=1 \sim N, i=0 \sim 6, \quad (4)$$

where $D_n^{(i)}$ and $S_n^{(i)}$ are independent random variables representing the distance and the shadowing between the MT in the 0th cell and the n th DA in the i th cell, respectively, γ denotes the path loss exponent. $\{S_n^{(i)} \mid n=1,2, \dots, N; i=0,1, \dots, 6\}$ are i.i.d. random variables with probability density function (PDF)

$$f_s(s) = \frac{1}{\sqrt{2\pi}\lambda\sigma_s s} \exp\left(-\frac{(\ln s)^2}{2\lambda^2\sigma_s^2}\right), s > 0, \quad (5)$$

where σ_s is the shadowing standard deviation and $\lambda = \frac{\ln 10}{10}$.

Since the number of interfering source is sufficiently large and interfering sources are independent with each other, the interference plus noise is assumed to be a complex Gaussian random vector as follows:

$$\mathbf{N} = \sum_{i=1}^6 \mathbf{H}^{(i)} \mathbf{x}^{(i)} + \mathbf{n}. \quad (6)$$

The variance of \mathcal{N} is derived by the Central Limit Theorem as

$$\text{Var}(\mathcal{N}) = \left[\sum_{i=1}^6 \sum_{n=1}^N [l_n^{(i)}]^2 P_n^{(i)} + \sigma_n^2 \right] \mathbf{I}_M \quad (7)$$

$$= \sigma^2 \mathbf{I}_M. \quad (8)$$

Therefore, (1) is rewritten as

$$\mathbf{y}^{(0)} = \mathbf{H}_w^{(0)} \mathbf{L}^{(0)} \mathbf{x}^{(0)} + \mathcal{N}. \quad (9)$$

3. Downlink capacity characterization

3.1 Problem formulation

Since the small-scale fading always varies fast but the large-scale fading usually varies quite slowly, we can regard $\mathbf{L}^{(0)}$ as a static parameter in calculating the downlink capacity. Thus, if the channel state information is only available at the receiver, the ergodic downlink capacity is calculated by taking expectation over the small-scale fading $\mathbf{H}_w^{(0)}$, which is expressed as (Telatar, 1999)

$$C = \mathbf{E}_{\mathbf{H}_w^{(0)}} \left[\log_2 \det \left(\mathbf{I}_M + \frac{1}{\sigma^2} (\mathbf{H}_w^{(0)} \mathbf{L}^{(0)}) \mathbf{P}^{(0)} (\mathbf{H}_w^{(0)} \mathbf{L}^{(0)})^H \right) \right], \quad (10)$$

where $P^{(0)} = \text{diag}\{P_1^{(0)}, P_2^{(0)}, \dots, P_N^{(0)}\}$ is the transmit power matrix of the DAs in the 0th cell. Unfortunately, it is quite difficult to get a more compact expression of the ergodic downlink capacity. Therefore, we propose the operation of “system scale-up” to study a simplified method to calculate the capacity as accurately as possible.

3.2 System scale-up

The basic idea of system scale-up is illustrated in Fig. 2. Assuming the proportion between the initial system and the scaled-up system to be t (a positive integer), we can summarize the characteristics of system scale-up as follows:

- Each DA with a single AE is scaled to a DA cluster with t AEs.
- The number of AEs equipped on the MT is increased from M to Mt .
- The system topology is not changed.
- The variance of \mathcal{N} is not changed.
- The large-scale channel fading is changed from $\mathbf{L}^{(i)}$ to $\mathbf{L}^{(i)t}$, we have

$$\mathbf{L}^{(i)t} = \mathbf{L}^{(i)} \otimes \mathbf{I}_t. \quad (11)$$

The small-scale fading is changed from $H_w^{(i)} \in \mathbb{C}^{M \times N}$ to $H_w^{(i)t} \in \mathbb{C}^{Mt \times Nt}$, $i = 0, 1, \dots, 6$, let $\mathbf{H}_{mn}^{(i)t} \in \mathbb{C}^{t \times t}$, $1 \leq m \leq M, 1 \leq n \leq N, i = 0, 1, \dots, 6$, we have

$$H_w^{(i)t} = \begin{bmatrix} \mathbf{H}_{11}^{(i)} & \dots & \mathbf{H}_{1N}^{(i)} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{11}^{(i)} & \dots & \mathbf{H}_{MN}^{(i)} \end{bmatrix}, i = 0, 1, \dots, 6. \quad (12)$$

$H_w^{(i)t}$ is also a matrix with i.i.d. zero-mean unit-variance circularly symmetric complex Gaussian entries, which is the same as $H_w^{(i)}$.

- The total power consumption is not changed. In detail, the transmit power of each DA in the initial system will be equally shared by the t AEs within a DA cluster in the scaled-up system. We can express the new transmit power matrix as

$$P_t^{(0)} = \frac{1}{t} \text{diag}\{P_1^{(0)}, P_2^{(0)}, \dots, P_N^{(0)}\} \otimes \mathbf{I}_t. \quad (13)$$

It is well known that the channel capacity of a MIMO system can be well approximated by a linear function of the minimum number of transmit and receive antennas (Telatar, 1999) as follows:

$$\mathbb{C} \approx \min(a, b) \times A, \quad (14)$$

where \mathbb{C} is the capacity of a MIMO channel with a transmit antennas and b receive antennas, A is a corresponding fixed parameter determined by the total transmit power constraint. If the number of transmit antennas and receive antennas increase from a to ta , from b to tb , respectively, the channel capacity is derived as

$$\hat{\mathbb{C}} \approx \min(ta, tb) \times A \approx t\mathbb{C}. \quad (15)$$

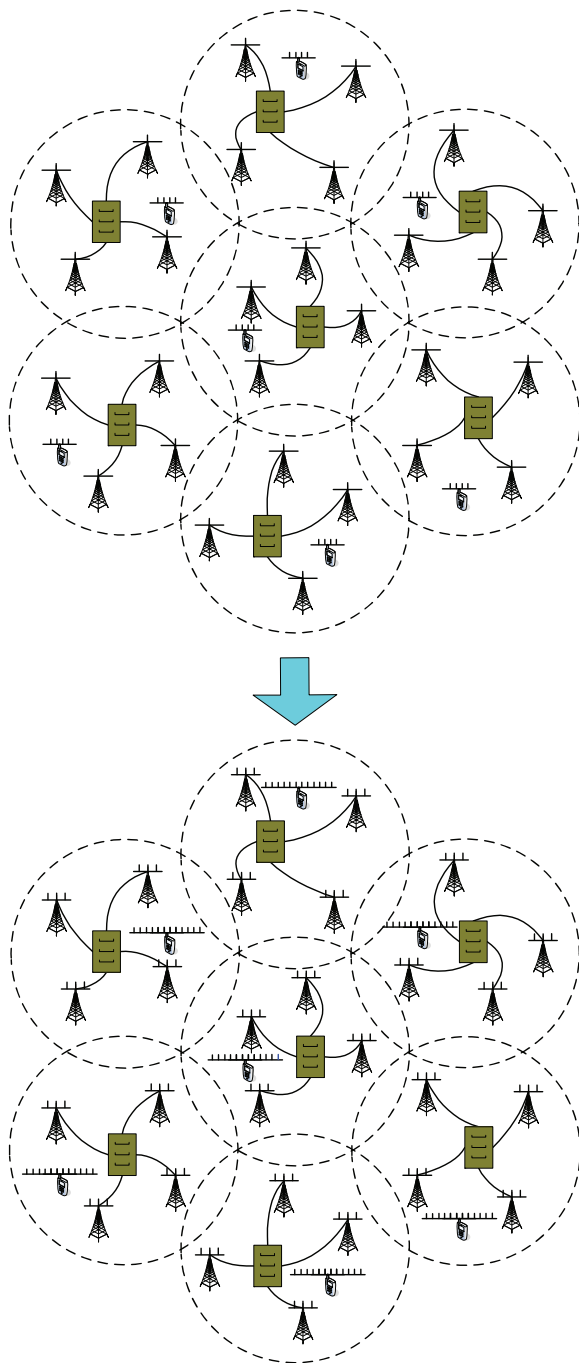


Fig. 2. Illustration of system scale-up.

Since the considered DAS is a special MIMO system, we directly hold that the system capacity scales linearly in the process of system scale-up, which can be partial testified by the following Theorem.

Theorem 1:

If the proportion between the initial system and the scaled-up system is t , an upper bound for the downlink capacity of the scaled-up system can be expressed as

$$C_t^{upper} = t \times M \log_2 \left(1 + \sum_{n=1}^N \frac{(I_n^{(0)})^2 P_n^{(0)}}{\sigma^2} \right). \quad (16)$$

Proof:

Based on (10), the capacity of the scaled-up system can be derived as

$$C_t = \mathbf{E}_{\mathbf{H}_w^{(0)t}} \left[\log_2 \det \left(\mathbf{I}_{Mt} + \frac{1}{\sigma^2} (\mathbf{H}_w^{(0)t} \mathbf{L}^{(0)t}) \mathbf{P}_t^{(0)} (\mathbf{H}_w^{(0)t} \mathbf{L}^{(0)t})^H \right) \right], \quad (17)$$

According to Hadamard Inequation, we have

$$C_t \leq \mathbf{E}_{\mathbf{H}_w^{(0)t}} \sum_{j=1}^{Mt} \log_2 \left(1 + \frac{1}{\sigma^2} \left[(\mathbf{H}_w^{(0)t} \mathbf{L}^{(0)t}) \mathbf{P}_t^{(0)} (\mathbf{H}_w^{(0)t} \mathbf{L}^{(0)t})^H \right]_{[jj]} \right), \quad (18)$$

Directly put \mathbf{E} inside \log_2 , according to Jensen Inequation, we have

$$C_t \leq \mathbf{E}_{\mathbf{H}_w^{(0)t}} \sum_{j=1}^{Mt} \log_2 \left(1 + \frac{1}{\sigma^2} \mathbf{E}_{\mathbf{H}_w^{(0)t}} \left[(\mathbf{H}_w^{(0)t} \mathbf{L}^{(0)t}) \mathbf{P}_t^{(0)} (\mathbf{H}_w^{(0)t} \mathbf{L}^{(0)t})^H \right]_{[jj]} \right), \quad (19)$$

Then, from (11), (12) and (13), we can further derive

$$C_t \leq t \times M \log_2 \left(1 + \sum_{n=1}^N \frac{(I_n^{(0)})^2 P_n^{(0)}}{\sigma^2} \right). \quad (20)$$

Thus,

$$C_t^{upper} = t \times M \log_2 \left(1 + \sum_{n=1}^N \frac{(I_n^{(0)})^2 P_n^{(0)}}{\sigma^2} \right). \quad (21)$$

■

Let

$$C_0 = M \log_2 \left(1 + \sum_{n=1}^N \frac{(I_n^{(0)})^2 P_n^{(0)}}{\sigma^2} \right). \quad (22)$$

From Theorem 1, we have

$$C_t^{upper} = t \times C_0. \quad (23)$$

It is observed that the upper bound of the system capacity scales linearly in the process of system scale-up.

3.3 Calculation of the downlink capacity

Based on the foregoing argument, we derive an approximation of the downlink ergodic capacity as

$$C \approx \frac{1}{t} \mathbf{E}_{\mathbf{H}_w^{(0)t}} \left\{ \log_2 \det \left[\mathbf{I}_{Mt} + \frac{1}{\sigma^2} \mathbf{H}_t^{(0)} \mathbf{P}_t^{(0)} \mathbf{H}_t^{(0)H} \right] \right\}, \quad (24)$$

where $\mathbf{H}_t^{(0)}$ is the channel matrix of the scaled-up system

$$\mathbf{H}_t^{(0)} = \begin{bmatrix} l_1^{(0)} \mathbf{H}_{11}^{(0)} & \dots & l_N^{(0)} \mathbf{H}_{1N}^{(0)} \\ \vdots & \ddots & \vdots \\ l_1^{(0)} \mathbf{H}_{M1}^{(0)} & \dots & l_N^{(0)} \mathbf{H}_{MN}^{(0)} \end{bmatrix}. \quad (25)$$

We rewrite (24) as

$$C \approx \frac{1}{t} \mathbf{E}_{\mathbf{H}_w^{(0)t}} \left\{ \log_2 \det \left[\mathbf{I}_{Mt} + \frac{1}{\sigma^2} \mathcal{H} \mathcal{H}^H \right] \right\}, \quad (26)$$

where $\mathcal{H} \in \mathbb{C}^{Mt \times Nt}$ and

$$H = \begin{bmatrix} l_1^{(0)} \sqrt{\frac{P_1^{(0)}}{t}} \mathbf{H}_{11}^{(0)} & \dots & l_N \sqrt{\frac{P_N^{(0)}}{t}} \mathbf{H}_{1N}^{(0)} \\ \vdots & \ddots & \vdots \\ l_1^{(0)} \sqrt{\frac{P_1^{(0)}}{t}} \mathbf{H}_{M1}^{(0)} & \dots & l_N \sqrt{\frac{P_N^{(0)}}{t}} \mathbf{H}_{MN}^{(0)} \end{bmatrix}. \quad (27)$$

Theorem 2:

The ergodic downlink capacity described in (10) can be accurately approximated as

$$C \approx \bar{C} = \sum_{n=1}^N \log_2 \left(1 + \frac{1}{\sigma^2} [l_n^{(0)}]^2 P_n^{(0)} W^{-1} M \right) + M \log_2(W) - M \log_2 e \left[1 - W^{-1} \right], \quad (28)$$

where W is the solution of the following equation

$$W = 1 + \sum_{n=1}^N \frac{[l_n^{(0)}]^2 P_n^{(0)}}{\sigma^2 + [l_n^{(0)}]^2 P_n^{(0)} W^{-1} M}. \quad (29)$$

Proof:

Let $v^t : [0, M) \times [0, N) \rightarrow \mathbb{R}$ be the variance profile function of matrix \mathcal{H} , which is given by

$$v^t(x, y) = t \cdot \text{Var}(\mathcal{H}(i, j)), \quad x \in \left[\frac{i-1}{t}, \frac{i}{t} \right); y \in \left[\frac{j-1}{t}, \frac{j}{t} \right),$$

where $\mathcal{H}(i, j)$ is the entry of matrix \mathcal{H} with index (i, j) . We can further find that as $t \rightarrow \infty$, $v_t(x, y)$ converges uniformly to a limiting bounded function $v(x, y)$, which is given by

$$v(x, y) = [I_n^{(0)}]^2 P_n^{(0)}, \quad x \in [0, M]; y \in [n-1, n]. \quad (30)$$

Therefore, the constraints of Theorem 2.53 in (Tulino & Verdu, 2004) are satisfied, we can derive the Shannon transform (Tulino & Verdu, 2004) of the asymptotic spectrum of $\mathcal{H}\mathcal{H}^H$ as

$$\begin{aligned} \mathcal{V}_{\mathcal{H}\mathcal{H}^H}(v) &= \lim_{t \rightarrow \infty} \frac{1}{t} \mathbf{E}_{\mathbf{H}_w^{(0)t}} [\log_2 \det(\mathbf{I} + v\mathcal{H}\mathcal{H}^H)] \\ &= E_{\mathbf{Y}} [\log_2 (1 + vE_{\mathbf{X}} [v(\mathbf{X}, \mathbf{Y})\Gamma_{\mathcal{H}\mathcal{H}^H}(\mathbf{X}, v) |\mathbf{Y}])]) \\ &\quad + E_{\mathbf{X}} [\log_2 (1 + vE_{\mathbf{Y}} [v(\mathbf{X}, \mathbf{Y})Y_{\mathcal{H}\mathcal{H}^H}(\mathbf{Y}, v) |\mathbf{X}])]) \\ &\quad - vE_{\mathbf{X}, \mathbf{Y}} [v(\mathbf{X}, \mathbf{Y})\Gamma_{\mathcal{H}\mathcal{H}^H}(\mathbf{X}, v)Y_{\mathcal{H}\mathcal{H}^H}(\mathbf{Y}, v)] \log_2 e, \end{aligned} \quad (31)$$

with $\Gamma_{\mathcal{H}\mathcal{H}^H}(\cdot, \cdot)$ and $Y_{\mathcal{H}\mathcal{H}^H}(\cdot, \cdot)$ satisfying the following equations

$$\Gamma_{\mathcal{H}\mathcal{H}^H}(x, v) = \frac{1}{1 + vE_{\mathbf{Y}} [v(x, \mathbf{Y})Y_{\mathcal{H}\mathcal{H}^H}(\mathbf{Y}, v)]}, \quad (32)$$

$$Y_{\mathcal{H}\mathcal{H}^H}(y, v) = \frac{1}{1 + vE_{\mathbf{X}} [v(\mathbf{X}, y)\Gamma_{\mathcal{H}\mathcal{H}^H}(\mathbf{X}, v)]}, \quad (33)$$

where \mathbf{X} and \mathbf{Y} represent independent random variables, which are uniform on $[0, M]$ and $[0, N)$, respectively, v is a parameter in Shannon transform. Given v , based on (30), we can observe that $\Gamma_{\mathcal{H}\mathcal{H}^H}(x, v)$ is constant on $x \in [0, M]$ and $Y_{\mathcal{H}\mathcal{H}^H}(y, v)$ is constant on $y \in [n-1, n)$. Thus, we define

$$\Gamma_{\mathcal{H}\mathcal{H}^H}(x, v) |_{x \in [0, M]} = W^{-1}, \quad (34)$$

$$Y_{\mathcal{H}\mathcal{H}^H}(y, v) |_{y \in [n-1, n)} = U_n^{-1}. \quad (35)$$

From (32) and (33), we have

$$W = 1 + v \sum_{n=1}^N [I_n^{(0)}]^2 P_n^{(0)} U_n^{-1}, \quad (36)$$

$$U_n = 1 + v [I_n^{(0)}]^2 P_n^{(0)} W^{-1} M, \quad 1 \leq n \leq N. \quad (37)$$

Assuming $v = \frac{1}{\sigma^2}$, we can further derive

$$C \approx \mathcal{V}_{\mathcal{H}\mathcal{H}^H} \left(\frac{1}{\sigma^2} \right) = \sum_{n=1}^N \log_2 \left(1 + \frac{1}{\sigma^2} [I_n^{(0)}]^2 P_n^{(0)} W^{-1} M \right) + M \log_2(W) - M \log_2 e \left[1 - W^{-1} \right]. \quad (38)$$

Moreover, from (36) and (37), W can be calculated by solving the following equation

$$W = 1 + \sum_{n=1}^N \frac{[I_n^{(0)}]^2 P_n^{(0)}}{\sigma^2 + [I_n^{(0)}]^2 P_n^{(0)} W^{-1} M}. \quad (39)$$

■

The unknown parameter W in Theorem 2 can be easily derived via an iterative method as presented in Table 1. The efficiency of the iterative algorithm will be demonstrated in Section 4.

<p><i>Initialization :</i></p> $W^0 = 1; \epsilon = 1.0 \times 10^{-6}.$ <p><i>Loopstep :</i></p> $W^l = 1 + \sum_{n=1}^N \frac{[l_n^{(0)}]^2 P_n^{(0)}}{\sigma^2 + [l_n^{(0)}]^2 P_n^{(0)} (W^{l-1})^{-1} M}$ <p><i>until</i> $\Delta = [W^l - W^{l-1}]^2 < \epsilon.$</p> <p><i>end.</i></p>
--

Table 1. The iterative method to calculate W .

4. Simulation results

In this section, Monte Carlo simulations are used to verify the validity of our analysis. The radius of a cell is assumed to be 1000m. The path loss exponent is set to be 4, the shadowing standard deviation is set to be 4 according to field measurement for microcell environment (Goldsmith & Greenstein, 1993), and the noise power σ_2^n is set to be -107dBm. The per distributed antenna power constraint takes value from -30dBm to 30dBm.

Without loss of generality, four different simulation setups are considered as follows:

- Case 1: $N = M = 4$, with randomly-selected system topology as shown in Fig. 3-A;
- Case 2: $N = M = 4$, with randomly-selected system topology as shown in Fig. 3-B;
- Case 3: $N = M = 8$, with randomly-selected system topology as shown in Fig. 3-C;
- Case 4: $N = M = 8$, with randomly-selected system topology as shown in Fig. 3-D;

Both analysis and simulation results of the ergodic downlink capacity for the four cases are presented in Fig. 4. It is observed that the two kinds of results are quite accordant with each other, which implies the high accuracy of the approximation in Theorem 2.

The total error covariance (Δ in Table 1) of the iterative method to calculate W is illustrated in Fig. 5. We can observe that 40 iteration steps are enough to make Δ be less than 1.0×10^{-6} .

In summary, we can conclude that the approximation is accurate and the iterative method is efficient.

5. Conclusions

In this chapter, the problem of characterizing the downlink capacity of a DAS with random antenna layout is addressed with the generalized assumptions: (a1) per distributed antenna power constraint, (a2) generalized mobile terminals equipped with multiple antennas, (a3) a multi-cell environment. Based on system scale-up, we derive a good approximation of the ergodic downlink capacity by adopting random matrix theory. We also propose an iterative method to calculate the unknown parameter in the approximation. The approximation is illustrated to be quite accurate and the iterative method is verified to be quite efficient by Monte Carlo simulations.

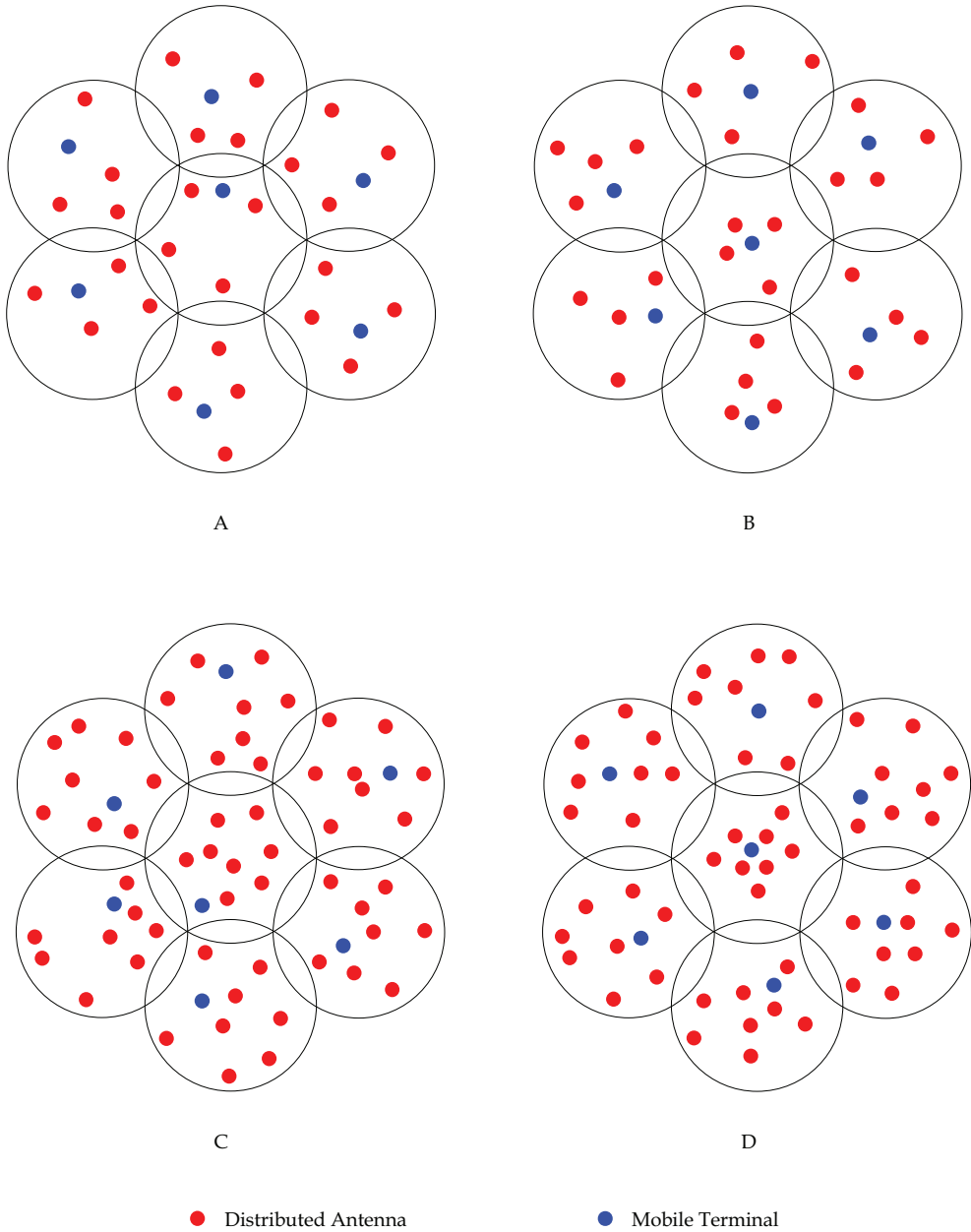


Fig. 3. Randomly-selected system topologies for simulations.

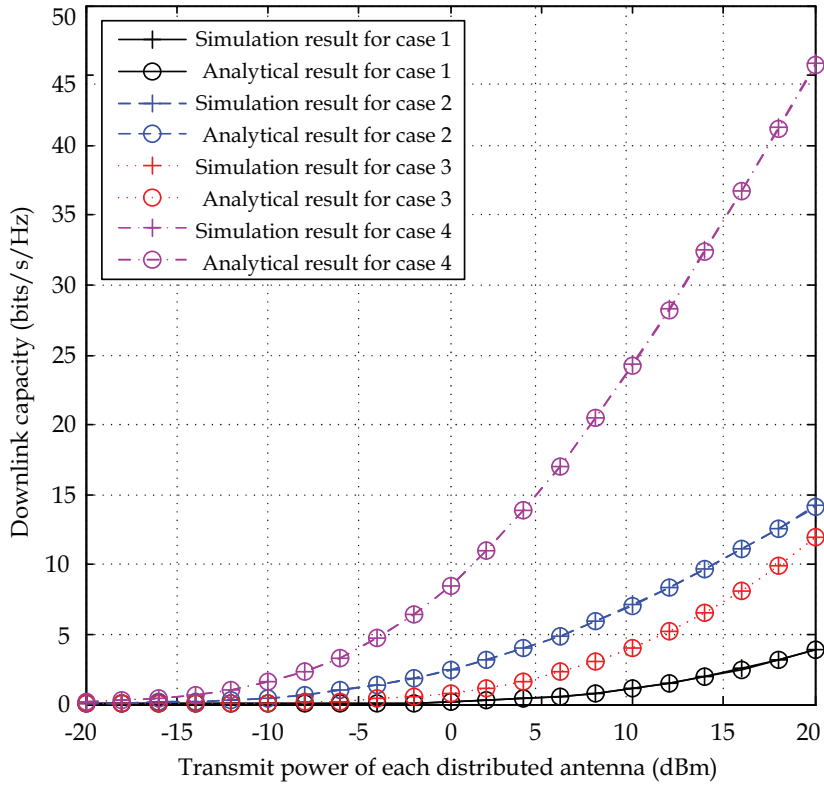


Fig. 4. Ergodic downlink capacity.

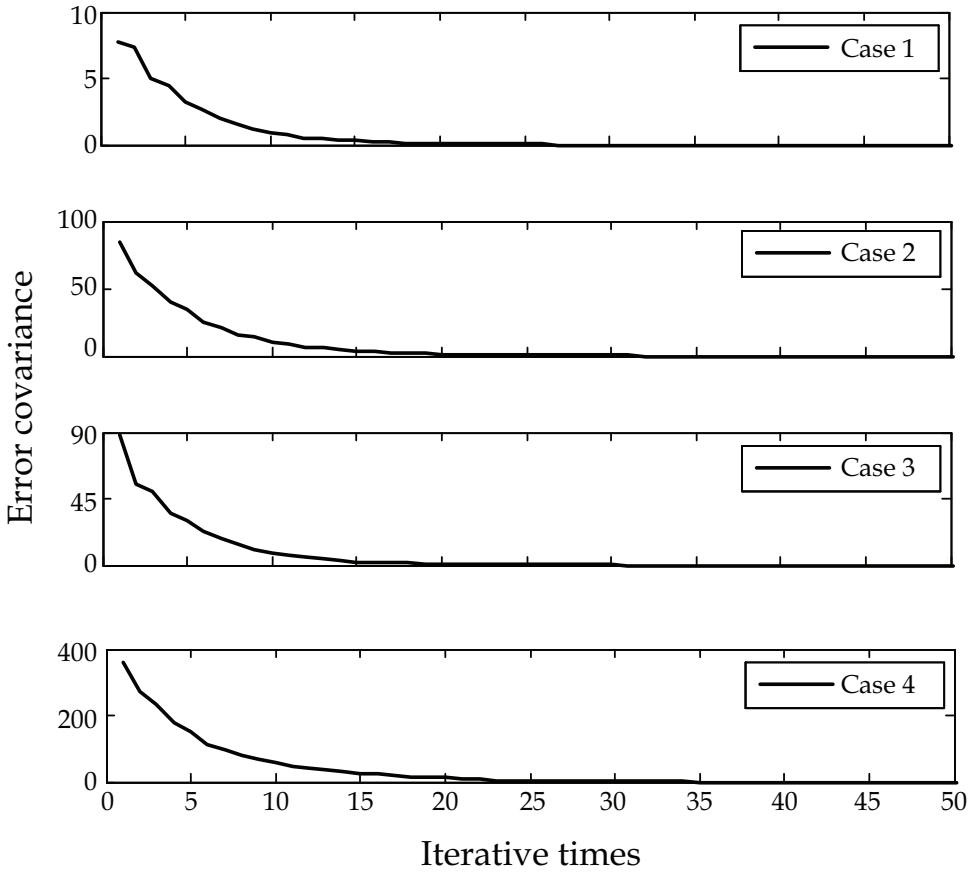


Fig. 5. Convergence performance of the iterative method.

6. References

- Choi, W. & Andrews, J. G. (2007). Downlink performance and capacity of distributed antenna systems in a multicell environment, *IEEE Trans. Wireless Commun.* Vol. 6(No. 1): 69–73.
- Feng, W., Li, Y., Gan, J., Zhou, S., Wang, J. & Xia, M. (2010). On the deployment of antenna elements in generalized multi-user distributed antenna systems, *Springer Mobile Networks and Applications* DOI. 10.1007/s11036-009-0214-1.
- Feng, W., Li, Y., Zhou, S., Wang, J. & Xia, M. (2009). Downlink capacity of distributed antenna systems in a multi-cell environment, *Proceedings of IEEE Wireless Commun. Networking Conf.*, pp. 1–5.
- Goldsmith, A. J. & Greenstein, L. J. (1993). Measurement-based model for predicting coverage areas of urban microcells, *IEEE J. on Selected Areas in Commun.* Vol. 11 (No. 7): 1013–1023.
- Hasegawa, R., Shirakabe, M., Esmailzadeh, R. & Nakagawa, M. (2003). Downlink performance of a cdma system with distributed base station, *Proceedings of IEEE Veh. Technol. Conf.*, pp. 882–886.
- Ni, Z. & Li, D. (2004). Effect of fading correlation on capacity of distributed mimo, *Proceedings of IEEE Personal, Indoor and Mobile Radio Commun. Conf.*, pp. 1637–1641.
- Roh, W. & Paulraj, A. (2002a). MIMO channel capacity for the distributed antenna systems, *Proceedings of IEEE Veh. Technol. Conf.*, pp. 706–709.
- Roh, W. & Paulraj, A. (2002b). Outage performance of the distributed antenna systems in a composite fading channel, *Proceedings of IEEE Veh. Technol. Conf.*, pp. 1520–1524.
- Saleh, A. A. M., Rustako, A. J. & Roman, R. S. (1987). Distributed antennas for indoor radio communications, *IEEE Trans. Commun.* Vol. 35(No. 12): 1245–1251.
- Telatar, E. (1999). Capacity of multi-antenna gaussian channels, *Eur. Trans. Telecomm.* Vol. 10(No. 6): 585–596.
- Tulino, A.M. & Verdú, S. (2004). *Random matrix theory and wireless communications*, The essence of knowledge, Princeton, New Jersey, USA.
- Xiao, L., Dai, L., Zhuang, H., Zhou, S. & Yao, Y. (2003). Information-theoretic capacity analysis in mimo distributed antenna systems, *Proceedings of IEEE Veh. Technol. Conf.*, pp. 779–782.
- Zhuang, H., Dai, L., Xiao, L. & Yao, Y. (2003). Spectral efficiency of distributed antenna systems with random antenna layout, *IET Electron. Lett.* Vol. 39(No. 6): 495–496.

Innovative Space-Time-Space Block Code for Next Generation Handheld Systems

Youssef Nasser and Jean-François H elard
*National Institute of Applied Sciences of Rennes
France*

1. Introduction

Broadcasting digital TV is currently an area of intensive development and standardisation activities. Actually, different groups are working on the standardisation problem. In Europe, the digital video broadcasting (DVB) consortium has adopted different standards for terrestrial (DVB-T) fixed reception, handheld (DVB-H) reception, satellite (DVB-S) reception as well as an hybrid reception like DVB-SH. In June 2008, the DVB-T2 was born extracting a lot of specifications from DVB-S2 and proposing some specifications for an eventual use of handheld reception. Now, we are working towards a second generation of DVB-SH called next generation handheld (NGH).

Technically, DVB-SH system provides an efficient and flexible mean of carrying broadcast services over an hybrid satellite and terrestrial infrastructure operating at frequencies below 3 GHz to a variety of portable, mobile and fixed terminals. Target terminals include handheld, vehicle-mounted, nomadic (e.g. laptops) and stationary terminals. The broadcast services encompass streaming services such as television, radio programs as well as download services enabling for example personal video recorder services. Typically, L-bands (1-2 GHz) and S-bands (2-4 GHz) are used for land mobile satellite (LMS) services. The DVB-SH system coverage is obtained by combining a satellite component (SATC) and, where necessary, a terrestrial component (TC) to ensure service continuity in areas where the satellite alone cannot provide the required quality of service (QoS). The SATC ensures wide area coverage while the TC provides cellular-type coverage. In the DVB-SH standard, two main physical layer configurations are supported. SH-A allows (but does not impose) a single frequency network (SFN) (Mattson, 2005) between the SATC and the TC, using the orthogonal frequency division multiplexing (OFDM) technique. SH-B supports a time division multiplexing (TDM) for the SATC and the OFDM technique for the TC.

The SFN presents great advantages by transmitting lower power at various sites throughout the coverage area. In an SFN, the different antennas transmit the same signal at the same moment on the same carrier frequency. The multi frequency network (MFN) allows however an optimization of the waveform and of the forward error correction (FEC) parameters according to the transmission environment. The existing SFN architectures are achieved in a single input single output system (SISO) since their deployment is very simple due to the use of one transmitting antenna by site. However, due to the increase of client services demand, it is desirable to deploy SFN with new multiple input multiple output (MIMO) techniques which ensure high spectrum efficiency as well as high diversity gain. In

DVB-T2, the multiple input single output (MISO) technique is adopted. In the future, the combination of MIMO and OFDM techniques is pursued as a potential candidate in standardisation and proposals. In the literature, there are few studies on the SFN with MIMO transmission. (Zhang et al, 2004) proposes a new SFN model to increase the diversity gain in MIMO SFN architecture. In (Kanbe et al, 2002), an array antenna receiver using a maximum ratio combining technique is proposed to improve the system performance of the SFN transmission. The lack of studies on this original idea motivates our work to extend the application of the MIMO-OFDM transmission to the SFN architecture and NGH systems. To the authors' knowledge, no contribution has been presented on the MIMO-OFDM technique for satellite-terrestrial NGH broadcasting systems.

The aim of this work is to propose efficient MIMO schemes for SFN and for combined satellite terrestrial (SATT) transmission for NGH broadcasting systems. In this proposition, two transmission scenarios are considered: the terrestrial transmission scheme and the hybrid satellite terrestrial transmission scheme. In terrestrial transmission scheme, we investigate the possibility of applying a space-time block code (STBC) encoder between the antennas of two sites in the SFN architecture. Secondly, a generalized framework is proposed for modeling the effect of unbalanced powers received from different transmitting antennas in MIMO-OFDM systems. This is a critical problem in SFN with mobile and portable reception. Then, we analyze and compare some of the most promising STBC schemes in the context of broadcasting for future terrestrial digital TV with equal and unequal received powers. Eventually, we propose a new 3D space-time-space (STS) block code for both environments. The use of a second space dimension will be justified as being particularly adapted and efficient in the case of SFN transmission but also in hybrid SATT transmission. The proposed code is based on a double level construction of ST codes resulting from the combination of two coding schemes. The first layer corresponds to an inter-cell ST coding while the second one corresponds to an intra-cell ST coding. In hybrid transmission, the adaptation of the proposed 3D MIMO scheme is more investigated due to the difference between the satellite channel link and the terrestrial links. In our contribution, we show that the 3D MIMO scheme is efficient in line of sight (LOS) situation (with respect to the satellite) but also in low, moderate and deep shadowing situations.

2. Functional areas addressed

The proposition treats the problem of MIMO block in a SFN scheme. The aim of this proposition is to improve the NGH system performance and to increase its spectral efficiency by using a distributed MIMO scheme. In this solution, two schemes are proposed: 4×2 and 2×2 antenna solutions. Moreover, the proposed solutions could be applied in terrestrial transmission scheme but also could be combined with a satellite component.

3. System model

3.1 System model in terrestrial SFN transmission

Classically, in SFN architectures, the different antennas transmit at the same moment the same signal on the same frequency. In this chapter, we analyze the application of a distributed MIMO communication scheme between the antennas located at different sites of the SFN architecture. Such a system could be implemented using M_T transmit antennas (Tx) by site as shown in Fig. 1. Without loss of generality, we will consider in our study the

transmission behavior of two neighboring cells using a total of $(2 \times M_T)$ Tx and M_R receive antennas (Rx). However, the proposed double layer scheme could be adapted to more than two neighboring cells.

From the reception viewpoint, one has to distinguish two cases: the open area and the gap area environments (Mattson, 2005) & (Zhang et al, 2004). Open areas correspond to the areas which benefit from an unobstructed view of transmitting antennas. Gap areas correspond to the areas where the direct signal is shadowed by obstacles. In the open area case, the transmitted signal is reflected from a large number of objects in the surroundings of the receiver. These signal components are received with independently time varying amplitudes and phases. In gap areas, there is no direct signal path and the transmitted signal might be easily lost due to the obstacles. To cope with gap areas, repeaters are usually implemented in those regions. The role of these repeaters, known as gap fillers, is to adequately amplify and retransmit the signal in the gap areas. Therefore, a receiver in the gap area could operate similarly as a receiver in the open area.

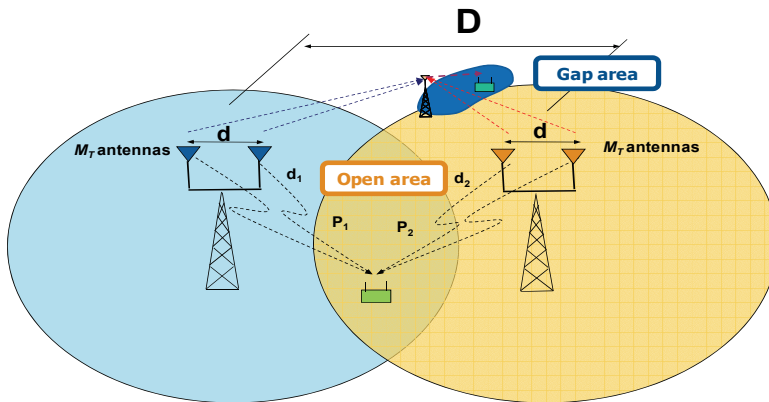


Fig. 1. SFN with unequal received powers

On the other hand, in order to make the SFN working properly the resulting total delay spread τ_{max} of the signals received from different antennas (located at different sites) must be less than the duration of the guard interval time inserted at the beginning of each OFDM symbol. Moreover, due to the path loss, the power received by each antenna is related to this resulting delay spread as it will be shown in next subsection. Thus, the selected MIMO scheme has to deal with those transmission conditions. As a starting point, let us assume that each site holds one antenna ($M_T=1$ in Fig. 1) and that the receiver receives signals from both antennas. Then, two receiving environments could be presented: the open area environment and the gap area environment.

3.1.1 Open area environment

In the case of an open area environment, the time offset between the signals received from each site antennas could be seen as a superposition of the time offset between transmitters' signals (the signal time delay between the transmitting antennas) and the signal time offset between each transmitter and the receiver. The first offset is generally negligible since the transmitters are synchronized with an ultra stable reference like the global positioning system (GPS). The second offset could be seen as follows. When the mobile terminal (MT)

moves within one cell, it receives signal from its own cell antenna but also from the neighboring cell antenna. Since the MT is not equidistant to both antennas, the signal received from each one will be delayed according to the position of the MT. This results into a delay $\Delta\tau$ between the two received signals from both antennas or equivalently between the channel impulse responses (CIR) between the transmitters and the receiver. The delays are directly related to the distances between the transmitters and the receiver and thus to the signal strength ratio at the receiver. Assuming an equal transmitted power P_0 at each antenna, the received power from the i^{th} antenna is:

$$P_i = \frac{P_0}{d_i^\alpha} \quad (1)$$

where d_i is the distance between the receiver and the i^{th} transmitter and α is the propagation constant which depends on the transmission environment.

The delay of each CIR between the i^{th} transmitter and the receiver is:

$$\tau_i = \frac{d_i}{c} \quad (2)$$

where c is the light celerity.

Without loss of generality, let us assume that the first transmitter site is the reference site. Substituting d_i from (2) in (1), the CIR delay of the i^{th} link (i.e. between the i^{th} transmitter and the receiver) with respect to the reference antenna can be expressed by:

$$\Delta\tau_i = \tau_i - \tau_1 = \left(10^{\frac{-\beta_i}{10\alpha}} - 1 \right) \frac{d_1}{c} \quad (3)$$

d_1 is the distance between the reference transmitter (first one) and the receiver. β_i is the received power difference (expressed in dB) between the signal received from the reference site and the signal received from the i^{th} transmitter. It is given by:

$$\beta_i[\text{dB}] = -10 \cdot \alpha \cdot \log_{10} \left(\frac{d_i}{d_1} \right) \quad (4)$$

In the sequel, we will assume that the power received from the reference antenna is equal to 0 dB and the distance d_i is greater than d_1 whatever i . It is a real situation where the MT is closer to its own cell antenna than other antennas. In this case, β_i is neither than the power attenuation factor between the i^{th} transmitter and the MT. As a consequence, the transmission model becomes equivalent to a system with unbalanced powers received from each site antennas. Fig. 2 shows an example of the relation between the power attenuation factor β and the timing offset between the CIR of the second link and the reference link in a system having two transmitting antennas located in two different sites.

If we now consider that the number of Tx in one site is greater than one i.e. $M_T > 1$, the choice of an adequate MIMO scheme should then be based on this imbalance in a SFN open area environment. Moreover, it should be adequate for inter-cell (i.e. between antennas signals of each site) and intra-cell (i.e. between antennas signals in each site) environments.

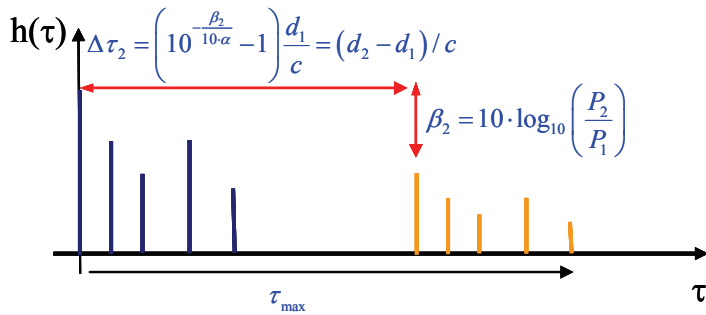


Fig. 2. Delays and powers in SFN

3.1.2 Gap area environment

In this case, the gap filler receives the signals from both antennas with a power difference according to equation (4). Then, it amplifies each one and retransmits them to the MT. Since the gap areas spread, in general, over a smaller region than that of the cell area, the delays between the signals received by the MT could be assumed equal to those between the signals received by the gap filler. Nevertheless, the signals received by the gap filler with a power imbalance of β_i could be re-amplified with different power gains which compensate this imbalance. As a consequence, the MT could receive the different signals with the same power. In other words, due to the gap filler, the delays between the received signals from both sites' antennas could be independent of the parameters β_i in the gap area environment. Thus, when designing a STBC for MIMO transmission, we should also consider this case of this terrestrial transmission scenario.

Based on this discussion on SFN transmission, it is clear that the STBC should be chosen adequately to cope with the open area and the gap area environments. This will be the subject of next sections where we propose a 3D STS code adapted for such situations.

3.1.3 3D MIMO scheme construction

Fig. 3 depicts the transmitter modules at each site. Information bits b_k are first channel encoded, randomly interleaved, and fed to a quadrature amplitude modulation (QAM) module. The SFN transmission system involving the two sites (described in Fig. 1) could therefore be seen as a double layer scheme in the space domain. The first layer is seen between the two sites separated by D km. The second layer is seen between the antennas separated by d meters within one site. For the first layer, a STBC scheme is applied between the 2 signals transmitted by each site antennas. In the second layer, we use a second STBC encoder for each subset of M_T signals transmitted from the same site. For the first layer (respectively the second layer), the STBC encoder takes L (respectively M) sets of data complex symbols and transforms them into a $(2, U)$ (respectively (M_T, V)) output matrix according to the STBC scheme. This output is then fed to $2 \times M_T$ OFDM modulators, each using N_c sub-carriers. In order to have a fair analysis and comparison between different STBC codes, the signal power at the output of the ST encoder is normalized by $2 \times M_T$.

The double layer encoding matrix is then described by:

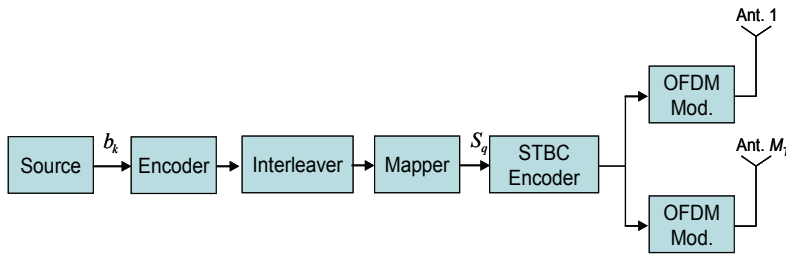


Fig. 3. MIMO-OFDM transmitter.

$$\mathbf{X}^{(1)} = \begin{pmatrix} \mathbf{X}_{11}^{(2)} & \cdots & \mathbf{X}_{1U}^{(2)} \\ \mathbf{X}_{21}^{(2)} & \cdots & \mathbf{X}_{2U}^{(2)} \end{pmatrix}$$

$$\mathbf{X}_{pq}^{(2)} = \begin{pmatrix} f_{pq,11}(s_1, \dots, s_M) & \cdots & f_{pq,1V}(s_1, \dots, s_M) \\ \vdots & \ddots & \vdots \\ f_{pq,M_T 1}(s_1, \dots, s_M) & \cdots & f_{pq,M_T V}(s_1, \dots, s_M) \end{pmatrix} \quad (5)$$

In (5), the superscript indicates the layer, $f_{pq,it}(s_1, \dots, s_M)$ is a function of the input complex symbols s_m and depends on the STBC encoder scheme. The time dimension of the resulting 3D code is equal to $U \times V$ and the resulting coding rate is $R = \frac{L \times M}{U \times V}$.

In order to simplify the transmission model, the double layer encoding matrix given in (5) will be represented by $\mathbf{X} = [x_{i,t}]$ where $x_{i,t}$ ($i=1, \dots, 2 \times M_T$; $t=1, \dots, U \times V$) is the output of the double layer STBC encoder on a given sub-carrier n . In other words, the layers construction is transparent from the transmission model viewpoint. Moreover, we set $Q=L \times M$ as the number of the complex symbols at the input of the double layer STBC encoder and we set $T=U \times V$ as the number of the corresponding output symbols. The ST coding rate is then $R=Q/T$.

The aim of this section is to judiciously build the proposed double layer 3D STS code so that the resulting MIMO scheme behaves efficiently in the SFN context for both aforementioned open and gap receiving environments. Then, we need to choose the adequate ST coding scheme, to apply on each layer of our 3D code.

First, we will consider the well-known orthogonal Alamouti ST coding scheme (Alamouti, 1998) for its robustness and simplicity. The Alamouti scheme provides full spatial diversity gain, no inter element interference (IEI), and requires low complexity maximum likelihood (ML) (Rupp et. al, 2004) receiver thanks to the orthogonality of its dispersion matrix. The Alamouti code has a rate R equal to one and its dispersion matrix is given by:

$$\mathbf{X} = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix} \quad (6)$$

For non-orthogonal schemes, we consider in this work the well-known space multiplexing (SM) scheme (Foschini, 1996). SM is designed to maximize the rate by transmitting symbols sequentially on different antennas. Its coding scheme is given by:

$$\mathbf{X} = [s_1 \quad s_2]^T \quad (7)$$

Finally, we consider the optimized Golden code (Belfiore et al, 2005) which is a full rate and fully diverse code. The Golden code is designed to maximize the rate such that the diversity gain is preserved for an increased signal constellation size. It is defined by:

$$\mathbf{X} = \frac{1}{\sqrt{5}} \begin{bmatrix} \beta(s_1 + \theta s_2) & \beta(s_3 + \theta s_4) \\ \mu \bar{\beta}(s_3 + \bar{\theta} s_4) & \bar{\beta}(s_1 + \bar{\theta} s_2) \end{bmatrix} \quad (8)$$

where $\theta = \frac{1+\sqrt{5}}{2}$, $\bar{\theta} = 1 - \theta$, $\alpha = 1 + j(1 - \theta)$, $\bar{\alpha} = 1 + j(1 - \bar{\theta})$.

To identify the most efficient ST code, the OFDM parameters are derived from those of the DVB-T standard (see Table 1). The spectral efficiencies 4 and 6 [b/s/Hz] are obtained for different ST schemes as shown in Table 2. In all simulations, we assume that two Rx antennas are used by the MT.

In the simulations results given hereafter, we separate the single layer case and the double layer case. For completeness point overview, we give first simulations results using a Rayleigh channel model in frequency domain i.e. we assume that the transmission from a transmitting antenna i to a receiving antenna j is achieved for each subcarrier n through a frequency non-selective Rayleigh fading channel. The use of the i.i.d. channel model is a first approach to justify our proposed 3D STS code. For this first step, the parameters β_i are chosen arbitrarily¹. In a second step, we will present the results with more realistic channel model like the COST 207 TU-6 channel model (COST, 1989). In this case, the results will be given for both, open and gap area environments.

FFT size	8K
Sampling frequency ($f_s=1/T_s$)	9.14 MHz
Guard interval (GI) duration	$1024 \times T_s = 112 \mu\text{s}$
Rate R_c of convolutional code	1/2, 2/3, 3/4
Polynomial code generator	(133,171) _o
Channel estimation	perfect
Constellation	16-QAM, 64-QAM, 256-QAM
Spectral Efficiencies	$\eta = 4$ and 6 [b/s/Hz]

Table 1. Simulations Parameters

Spectral Efficiency	ST scheme	ST rate R	Constellation	R_c
$\eta=4$ [bit/Sec/Hz]	Alamouti	1	64-QAM	2/3
	SM	2	16-QAM	1/2
	Golden	2	16-QAM	1/2
	3D code	2	16-QAM	1/2
$\eta=6$ [bit/Sec/Hz]	Alamouti	1	256-QAM	3/4
	SM	2	64-QAM	1/2
	Golden	2	64-QAM	1/2
	3D code	2	64-QAM	1/2

Table 2. Different MIMO schemes and efficiencies

¹ Since we model the channel in frequency domain, there is no CIR and hence no CIR delays in this case.

a. Single Layer case: inter-cell ST coding

The received signal at the input of the MT could be written as:

$$\mathbf{y} = \mathbf{G}\mathbf{B}\mathbf{F}\mathbf{s} + \mathbf{w} = \mathbf{G}_{\text{eq}}\mathbf{s} + \mathbf{w} \quad (9)$$

where the matrix \mathbf{G} is composed of blocks $\mathbf{G}_{j,i}$ ($j=1,\dots,M_R; i=1,\dots,2M_T$) each having $(2T, 2T)$ elements, reflecting the channel coefficients (Khalighi et al., 2006) & (Nasser et al., 2008). \mathbf{B} is the matrix reflecting the powers received from each antenna. \mathbf{F} is composed of $2M_T$ blocks of $2T$ rows each i.e. the data transmitted on each antenna are gathered in one block having $2T$ rows and $2Q$ columns according to the ST coding scheme. \mathbf{G}_{eq} is the equivalent channel matrix between \mathbf{s} and \mathbf{y} . It is assumed to be known perfectly at the receiving side.

The optimal receiver is a ML (Rupp et al., 2004) receiver whose complexity increases exponentially with the number of antennas and the constellation size. In the case of orthogonal STBC (OSTBC), the optimal receiver is simply made of a concatenation of ST decoder and channel decoder modules. However, in the case of non-orthogonal STBC (NO-STBC) schemes, there is an ICI at the receiving side. The optimal receiver becomes more complex since it requires joint ST and channel decoding operations. Moreover, it requires large memory to store the different points of the trellis. In our work, we use a sub-optimal solution based on an iterative receiver where the ST detector and channel decoder exchange extrinsic information in order to enhance soft information metrics. The iterative detector shown in Fig. 4 is composed of a parallel interference canceller (PIC), a demapper which consists in computing the soft information of the transmitted bits, i.e. a log likelihood ratio (LLR) computation (Tosato & Bisaglia, 2002), a soft-input soft-output (SISO) decoder (Hagenauer & Hoehner, 1989), and a soft mapper.

At the first iteration, the demapper takes the estimated symbols $\hat{\mathbf{s}}$, the knowledge of the channel \mathbf{G}_{eq} and of the noise variance, and computes the LLR values of each of the coded bits transmitted per channel use. The estimated symbols $\hat{\mathbf{s}}$ are obtained via minimum mean square error (MMSE) filtering according to:

$$\hat{\mathbf{s}}_p^{(1)} = \mathbf{g}_p^{\text{tr}} (\mathbf{G}_{\text{eq}} \cdot \mathbf{G}_{\text{eq}}^{\text{tr}} + \sigma_w^2 \mathbf{I})^{-1} \mathbf{y} \quad (10)$$

where \mathbf{g}_p^{tr} of dimension $(2M_R T, 1)$ is the p^{th} column of \mathbf{G}_{eq} ($1 \leq p \leq 2Q$). $\hat{\mathbf{s}}_p^{(1)}$ is the estimation of the real part (p odd) or imaginary part (p even) of s_q ($1 \leq q \leq Q$). Once the estimation of the different symbols s_q is achieved by the soft mapper at the first iteration, we use this estimation for the next iterations process.

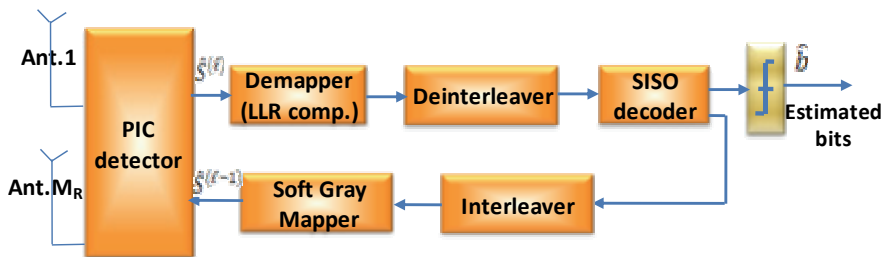


Fig. 4. Iterative receiver structure

From the second iteration, we perform PIC operation followed by a simple inverse filtering (instead of MMSE filtering at the first iteration). This block offers enhanced LLR values to be fed to the SISO decoder by suppressing the spatial interference. This interference cancellation is described by:

$$\begin{aligned}\hat{\mathbf{y}}_p &= \mathbf{y} - \mathbf{G}_{\text{eq},p} \tilde{\mathbf{s}}_p^{(1)} \\ \hat{\mathbf{s}}_p^{(2)} &= \frac{1}{\mathbf{g}_p^{\text{tr}} \mathbf{g}_p} \mathbf{g}_p^{\text{tr}} \hat{\mathbf{y}}_p\end{aligned}\quad (9)$$

where $\mathbf{G}_{\text{eq},p}$ of dimension $(2M_R T, 2Q-1)$ is the matrix $\mathbf{G}_{\text{eq},p}$ with its p^{th} column removed, $\tilde{\mathbf{s}}_p^{(1)}$ of dimension $(2Q-1, 1)$ is the vector $\tilde{\mathbf{s}}$ estimated by the soft mapper with its p^{th} entry removed. In our proposition, we consider a sub-optimal iterative detector for non-orthogonal schemes in order to cancel the IEI. The iterative process used here converges after 3 iterations (Nasser et al., May 2008). Therefore, all the results given thereafter are obtained after 3 iterations.

In the case of single layer reception, we have one antenna by site. Then, the second layer matrix $\mathbf{X}^{(2)}$ in (5) resumes to one element. The multiple input component of the MIMO scheme is then only obtained by the single antenna in each site ($M_T=1$). Due to the mobility, the MT is assumed to occupy different locations and the first layer ST scheme must be efficient face to unequal received powers. For equal received powers, we assume that the powers of the matrix \mathbf{B} in (9) are equal to 0 dB.

Fig. 5 gives the required E_b/N_0 to obtain a $\text{BER}=10^{-4}$ for a spectral efficiency $\eta=4$ [b/s/Hz]. Moreover, since we have one Tx antenna by site, we set $\beta_1=0$ dB and we change $\beta=\beta_2$. As expected, this figure shows that the Golden code presents the best performance when the Rx receives the same power from both sites (i.e. $\beta_1=\beta_2=0$ dB). When β_2 decreases, the Alamouti scheme is very efficient and presents a maximum loss of only 3 dB in terms of required E_b/N_0 with respect to equal received powers case. Indeed, for very small values of β , the

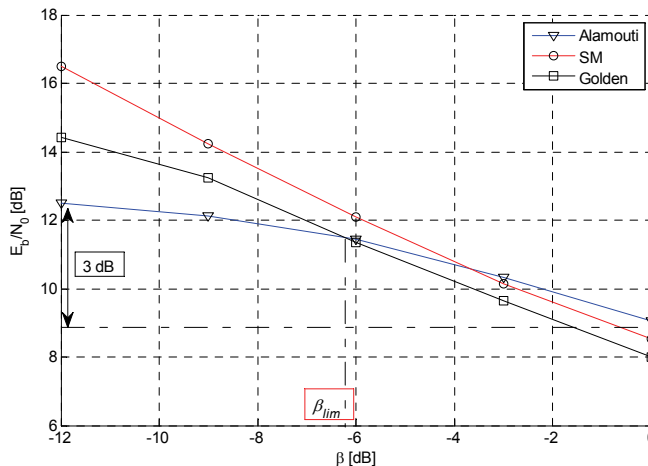


Fig. 5. Required E_b/N_0 to obtain a $\text{BER}=10^{-4}$, single layer case, $\eta=4$ [b/s/Hz]

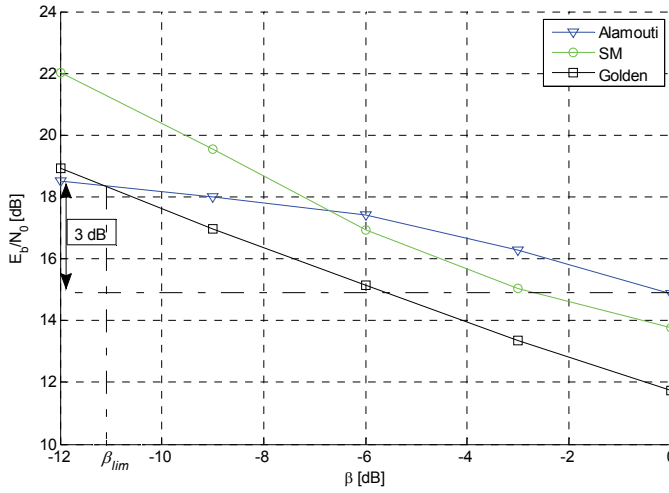


Fig. 6. Required E_b/N_0 to obtain a $BER=10^{-4}$, single layer case, $\eta=4$ [b/s/Hz]

transmission scenario becomes equivalent to a transmission scenario with one transmitting antenna. In this figure, the value $\beta_{lim} = -6.2$ dB presents the power imbalance limit where the Alamouti and the Golden code schemes have the same performance at a $BER=10^{-4}$.

The same kind of results, shown in Fig. 6, are observed for a spectral efficiency $\eta=6$ [b/s/Hz]. The Alamouti scheme is more efficient than the Golden code when the power imbalance parameter β becomes less than a given limit value $\beta_{lim} = -11.2$ dB.

b. Double Layer case: intra-cell ST coding

Considering the whole double layer space domain construction, one ST coding scheme has to be assigned to each layer of the proposed system. The resulting 3D STS code should be efficient for both environments in SFN architecture. We propose to construct the first layer, i.e. the inter-cell coding, with Alamouti scheme since it is the most resistant for the unequal received powers case. In a complementary way, we propose to construct the second layer, i.e. the intra-cell coding, with the Golden code since it offers the best results in the case of equal received powers. After combination of the two space layers with time dimension, (5) yields:

$$X = \frac{1}{\sqrt{5}} \begin{pmatrix} \alpha(s_1 + \theta s_2) & \alpha(s_3 + \theta s_4) & \alpha(s_5 + \theta s_6) & \alpha(s_7 + \theta s_8) \\ j\bar{\alpha}(s_3 + \bar{\theta} s_4) & \bar{\alpha}(s_1 + \bar{\theta} s_2) & j\bar{\alpha}(s_7 + \bar{\theta} s_8) & \bar{\alpha}(s_5 + \bar{\theta} s_6) \\ -\alpha^*(s_5^* + \theta^* s_6^*) & -\alpha^*(s_7^* + \theta^* s_8^*) & \alpha^*(s_1^* + \theta^* s_2^*) & \alpha^*(s_3^* + \theta^* s_4^*) \\ j\alpha^*(s_7^* + \bar{\theta}^* s_8^*) & -\alpha^*(s_5^* + \bar{\theta}^* s_6^*) & -j\alpha^*(s_3^* + \bar{\theta}^* s_4^*) & \alpha^*(s_1^* + \bar{\theta}^* s_2^*) \end{pmatrix} \quad (12)$$

where $\theta = \frac{1+j\sqrt{5}}{2}$, $\bar{\theta} = 1 - \theta$, $\alpha = 1 + j(1 - \theta)$, $\bar{\alpha} = 1 + j(1 - \bar{\theta})$.

3.1.4 Simulation results in open area environment

In an open area environment, the MT is in an unobstructed region with respect to each site antennas. Since the distance d between the transmitting antennas in one site is negligible

with respect to the distance D (Fig. 1), the power attenuation factors in the case of our 3D code are such that $\beta_1=\beta_2=0$ dB and $\beta=\beta_3=\beta_4$.

Fig. 7 shows the results in terms of required E_b/N_0 to obtain a BER equal to 10^{-4} for different values of β and 3 STBC schemes i.e. our proposed 3D code scheme, the single layer Alamouti and the Golden code schemes assuming Rayleigh i.i.d frequency channel coefficients. In this figure, the value β corresponds to β_2 for the single layer case and to $\beta=\beta_3=\beta_4$ for our 3D code. Fig. 7 shows that the proposed scheme presents the best performance whatever the spectral efficiency and the factor β are. Indeed, it is optimized for SFN systems and unbalanced received powers. For $\beta=-12$ dB, the proposed 3D code offers a gain equal to 1.8 dB (respectively 3 dB) with respect to the Alamouti scheme for a spectral efficiency $\eta=4$ [b/s/Hz] (resp. $\eta=6$ [b/s/Hz]). This gain is greater when it is compared to the Golden code. Moreover, the maximum loss of our code due to unbalanced received powers is equal to 3 dB in terms of E_b/N_0 . This means that it leads to a powerful code for SFN systems.

In a MIMO COST 207 TU-6 channel model, we assume that the MT is moving with a velocity of 10 km/h and the distance d_1 between the receiver and the reference antenna is equal to 5 km. The CIRs between different transmitters and the MT are delayed according to (3).

Fig. 8 gives the same kind of results of those given in Fig. 7. Once again, these results highlight the superiority of the proposed 3D code in real channel models whatever the spectral efficiency and the factor β i.e. the 3D code outperforms the others schemes in all cases. The gain could reach 1.5 dB for a spectral efficiency $\eta=4$ [b/s/Hz] and 3.1 dB for a spectral efficiency $\eta=6$ [b/s/Hz].

Fig. 9 evaluates the robustness of the different schemes to the MT velocity. We assume that the MT is moving within one cell with a velocity of 10 km/h and 60 km/h respectively. We show in this figure that the Alamouti scheme is very robust to the MT velocity. The degradation of the Golden code might reach 1 dB in terms of required E_b/N_0 to reach a BER= 10^{-4} . The degradation of the proposed 3D code due to the MT velocity and hence to the Doppler effect is of about 0.2 dB only.

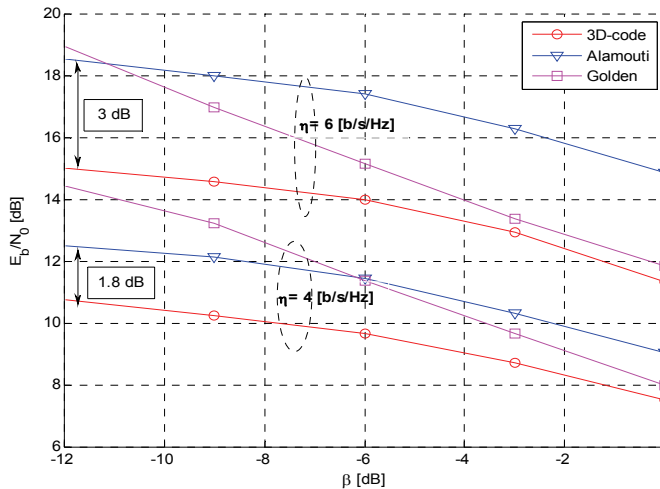


Fig. 7. Required E_b/N_0 to obtain a BER= 10^{-4} , double layer case, $\eta=4$ [b/s/Hz], $\eta=6$ [b/s/Hz], Rayleigh channel

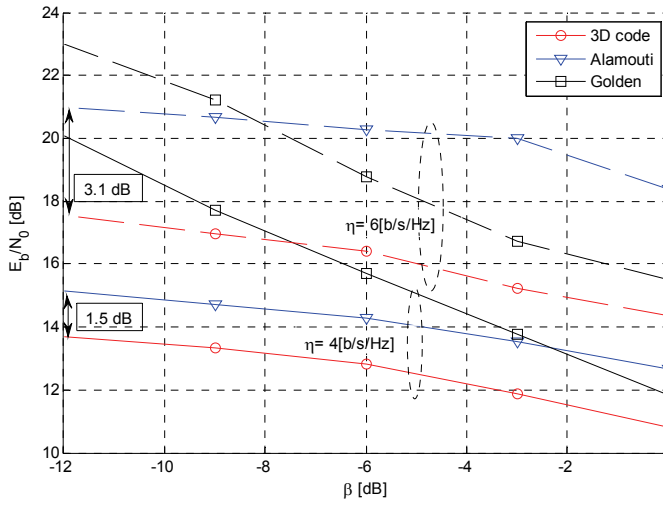


Fig. 8. Required E_b/N_0 to obtain a BER= 10^{-4} , double layer case, $\eta=4$ [b/s/Hz], $\eta=6$ [b/s/Hz], TU-6 channel

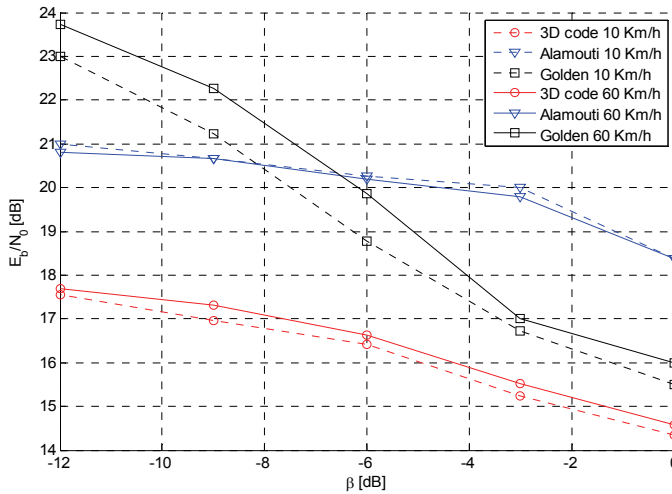


Fig. 9. Required E_b/N_0 to obtain a BER= 10^{-4} , $\eta=6$ [b/s/Hz], TU-6 channel, different values of MT velocity

3.1.5 Simulation results in gap area environment

In a gap area environment, the MT is in obstruction with respect to each site antennas. In this case, the gap filler receiving antennas become at the same situation of those of the MT in the open area environment i.e. a power imbalance is observed at the receiving side and it is related to the CIR delays by equation (4). However, due to the gap filler amplification, the power received by the MT in a gap area could be independent of these delays.

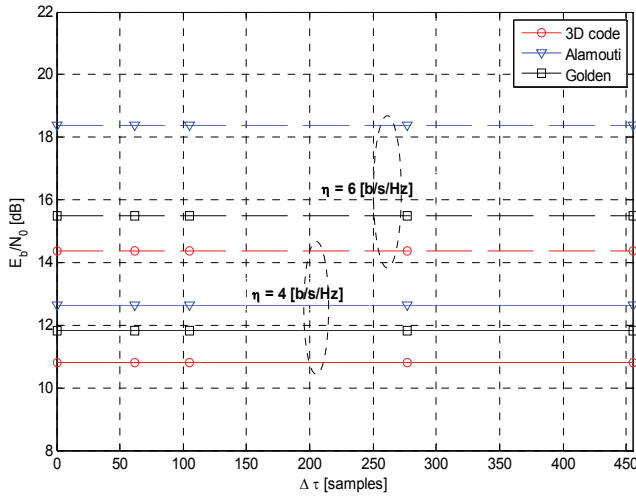


Fig. 10. Required E_b/N_0 to obtain a $BER=10^{-4}$, $\eta=4$ [b/s/Hz], $\eta=6$ [b/s/Hz], TU-6 channel, gap area environment.

In Fig. 10, we give the required E_b/N_0 that a MT needs in a gap area to obtain a $BER = 10^{-4}$ with respect to the CIR delays $\Delta\tau$ observed at the gap filler receivers. As expected, we show in this figure that the results are independent of these delays since they are smaller than the guard interval durations (GI= 1024 samples). In other words, as these delays are less than the guard interval duration, they produce only a phase rotation which is corrected by the equalizer in the frequency domain. The power imbalance is already corrected by the gap filler amplification.

3.2 System model in hybrid satellite terrestrial transmission

For hybrid SATT transmission, we propose to apply the MIMO scheme between the terrestrial and satellite sites as described in Fig. 11. Due to the links model difference, i.e. satellite link and terrestrial link, the proposed code has to cope with different transmission scenarios. More precisely, the MIMO scheme has to be efficient in the LOS region but also in shadowing regions (moderate and deep) with respect to the satellite antennas. In order to achieve that, we propose again to use the 3D MIMO scheme for such situations. The first layer corresponds to the inter-cell ST coding, i.e. between satellite and terrestrial antennas, while the second corresponds to the intra-cell ST coding, i.e. between the antennas of the same site. For the satellite links, we have considered the land mobile satellite (LMS) (Murr et al., 1995) adopted in DVB-SH (ETSI, 2008) and described by Fontan (Fontan et al., 2001), (Loo, 1985) & (Fontan et al., 1998). The LMS channel is modeled by Markov chain with three states. The state S1 corresponds to the LOS situation, while S2 and S3 correspond respectively to the moderate and deep shadowing situations. Generally speaking, the LMS channel in each state follows a Loo distribution (Loo, 1985). The latter is a Rice distribution where its mean follows a log-normal distribution having a mean μ and a standard deviation Σ . Table 3 shows that the different states of the Markov chain depend on the elevation angles and that each state has its specified mean and standard deviation. The parameter MP in this table reflects the multipath component power in the Rice distribution.

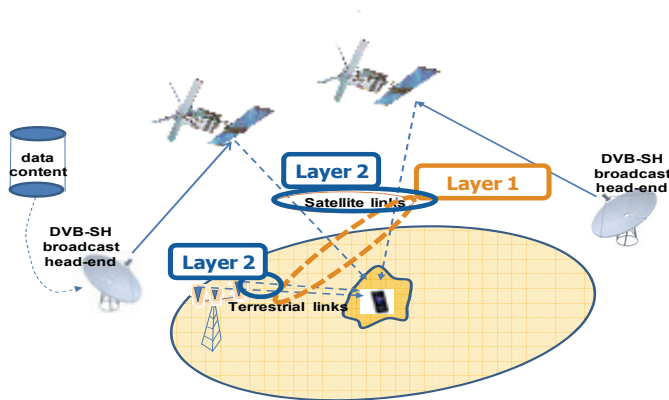


Fig. 11. Layered STS 3D code using SATT transmission scheme

3.2.1 First layer construction: SATT coding

In order to construct the first layer, we consider the same method as done for terrestrial transmission. First, we will construct the first layer using the well-known MIMO schemes, i.e. Alamouti and Golden codes. Second, due to the mobility, the MT is assumed to occupy different locations over a sufficient long route. Then, the first layer ST scheme must be efficient face to shadowing during its trajectory. Recall that the moderate and deep shadowing are dependent of the elevation angle. For example, in Table 3, the moderate shadowing for an elevation angle of 30° corresponds to a mean value $\mu = -4.7$ dB and the deep shadowing corresponds to a mean value equal to -7 dB. It is clear from Table 3 that for an elevation angle equal to 30° , the system presents the highest signal power level since the moderate and deep shadowing are relatively acceptable comparing to other elevation angles θ . In the sequel, we will present first the results obtained with an elevation angle $\theta = 30^\circ$ and $\theta = 50^\circ$ and for the various spectral efficiencies using an Alamouti and Golden code scheme at the first layer. Fig. 12 shows the required E_b/N_0 to obtain a BER equal to 10^{-4} for a spectral efficiency $\eta = 2, 4$ and 6 b/s/Hz. As expected, we conclude from these results that for low spectral efficiency, i.e. $\eta = 2$, the Alamouti scheme outperforms the Golden scheme. However, for a spectral efficiency $\eta = 4$ and $\eta = 6$, the conclusion on the best performance is not immediate. It depends on the elevation angle and hence on the shadowing level. For high shadowing level (see Table 3, $\theta = 50^\circ$), the Alamouti code presents almost better

Elevation	S1: LOS			S2: Interm. Shadowing			S3: Deep Shadowing		
	μ	Σ	MP	μ	Σ	MP	μ	Σ	MP
10°	-0.1	0.5	-19	-8.7	3	-12	-12.1	6	-25
30°	-0.5	1	-15	-4.7	1.5	-19	-7	3	-20
50°	-0.5	1	-17	-6.5	2.5	-17	-14	2.5	-20
70°	-0.2	0.5	-15	-6.0	2.1	-17	-11.5	2	-20

Table 3. Average Loo model parameters in dB for various angles and suburban area (measurement results given in (Fontan et al., 1985))

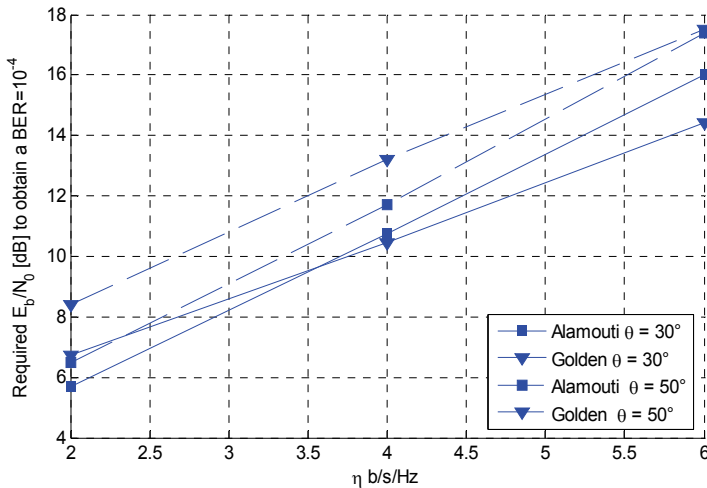


Fig. 12. Required E_b/N_0 to obtain a $BER=10^{-4}$, single layer case

performance. In summary, the Golden code scheme outperforms the Alamouti scheme only for high spectral efficiency and relatively low shadowing levels ($\theta = 30^\circ$). This confirms our results in terrestrial transmission where the performance of the MIMO scheme depends on the power imbalance between the two signals received from each site.

3.2.2 Second layer construction: intra-site coding

Considering the whole layers' construction (i.e. $M_T > 1$), one ST coding scheme has to be assigned to the SATT coding and another ST coding scheme has to be assigned to the intra-site coding. The resulting layered ST coding should be efficient for low, moderate and deep shadowing levels. Considering the results and conclusions obtained in previous sub-section, we propose to construct the SATT layer with Alamouti scheme, since it is the most resistant for the deep shadowing levels. In a complementary way, we propose to construct the second layer with the Golden code since it offers the best results in the case of relatively low shadowing levels.

Fig. 13 shows the results in terms of required E_b/N_0 to obtain a BER equal to 10^{-4} for the various elevation angles, two spectral efficiencies $\eta = 2$ b/s/Hz and $\eta = 6$ b/s/Hz and the three considered codes i.e. our proposed 3D scheme, the single layer Alamouti scheme and the single layer Golden scheme. The results obtained in this figure show that the proposed 3D scheme outperforms the other schemes whatever the elevation angle and the spectral efficiency are. Moreover, as expected, the best performance is obtained for an elevation angle $\theta = 30^\circ$. The gain of the 3D code compared to the Alamouti scheme is about 1 dB for $\eta = 2$ b/s/Hz and can reach 4 dB for $\eta = 6$ b/s/Hz. The conclusions of Fig. 13 are confirmed in Fig. 14 for $\eta = 4$ b/s/Hz. This means that the 3D code leads to a powerful code for next DVB-NGH systems.

3.3 Conclusions

In this work, we have presented a full rate full diversity 3D code, a promising candidate for next generation broadcast technologies. It is constructed using two layers: the first layer

using Alamouti code and the second layer using Golden code. We showed that our proposed scheme is very efficient to cope with low, moderate and deep shadowing levels as well as various elevation angles. The proposed scheme is fully compatible with SFN and hybrid SATT scheme. It is then a very promising candidate for the broadcasting of the future terrestrial digital TV through NGH structures.

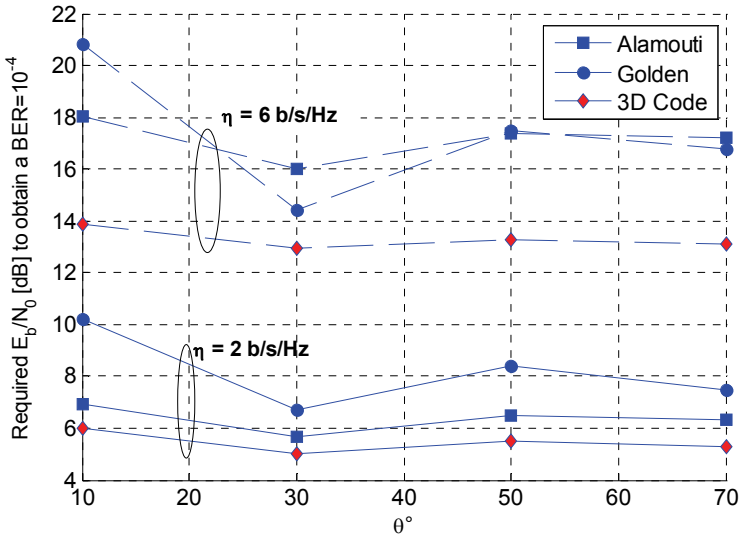


Fig. 13. Required E_b/N_0 to obtain a BER= 10^{-4} , double layer construction, η variable

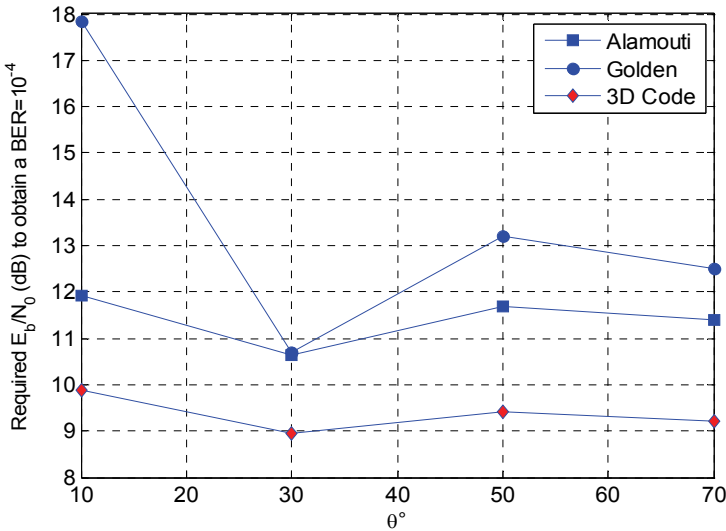


Fig. 14. Required E_b/N_0 to obtain a BER= 10^{-4} , double layer construction, $\eta=4 \text{ b/s/Hz}$

4. References

- Mattson A. (2005), Single frequency networks in DTV, *IEEE Trans. on Broadcasting*, Vol. 51, Issue 4, Dec. 2005, pp. 413-422, ISSN : 0018-9316.
- Zhang L., Gui L., Qiao Y., and Zhang W. (2004), Obtaining diversity gain for DTV by using MIMO structure in SFN, *IEEE Trans. on broadcasting*, Vol. 50, No. 1, March 2004, 83-90, ISSN: 0018-9316.
- Kanbe Y., Itami M., Itoh K., and Aghvami A. (2002), Reception of an OFDM signal with an array antenna in a SFN environment, *Proc. of IEEE Personal Indoor and Mobile Radio Communications*, Vol. 3, 1310-1315, ISBN: 0-7803-7589-0, Sept. 2002.
- Alamouti, S.M. (1998), A simple transmit diversity technique for wireless communications, *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 8, Oct. 1998, 1451-1458, ISSN: 0733-8716.
- Rupp M., Gritsch G., Weinrichter H. (2004), Approximate ML detection for MIMO systems with very low complexity, *Proc. of the International conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 809-812, ISBN: 0-7803-8484-9, May 2004.
- Foschini G. J. (1996), Layered space-time architecture for wireless communication in a fading environment when using multi-element antenna, *Bell Labs Tech. Journal*, Vol. 1, no. 2, 41-59.
- Belfiore, J.-C., Rekaya G., & Viterbo E. (2005). The golden code: a 2×2 full-rate space-time code with non vanishing determinants, *IEEE Transactions in Information Theory*, Vol. 51, No. 4, April 2005, 1432-1436, ISSN : 0018-9448.
- COST (1989), *COST 207 Report*, Digital Land Mobile Radio Communications, Commission of European Communities, Directorate General, Telecommunications Information Industries and Innovation, Luxemburg.
- Khalighi M. A., Hélar J.-F., and Bourennane S. (2006), Contrasting Orthogonal and non orthogonal space-time schemes for perfectly-known and estimated MIMO channels, *Proc. of IEEE Int. Conf. on Communications systems*, 1-5, ISBN: 1-4244-0411-8, Oct. 2006, Singapore.
- Nasser, Y.; Helard, J.-F. & Crussiere, M. (2008). System Level Evaluation of Innovative Coded MIMO-OFDM Systems for Broadcasting Digital TV. *International Journal of Digital Multimedia Broadcasting*, Vol. 2008, pages 12, doi:10.1155/2008/359206.
- Nasser Y., Hélar J.-F., Crussiere M., and Pasquero O. (2008), Efficient MIMO-OFDM schemes for future terrestrial digital TV with unequal received powers, *Proc. of IEEE International Communications Conference, 2021 - 2027*, ISBN: 978-1-4244-2075-9, June 2008, Beijing, China.
- Tosato F., and Bisaglia P. (2002), Simplified Soft-Output Demapper for Binary Interleaved COFDM with Application to HIPERLAN/2, *Proc IEEE Int. Conf. on Communications*, pp. 664-668, ISBN: 0-7803-7400-2, June 2002.
- Hagenauer J., and Hoeher P. (1989), A Viterbi algorithm with soft-decision outputs and its applications, *Proc. of IEEE Global Telecommunications Conf.*, pp. 1680-1686, Nov. 1989, Dallas, USA.
- Murr F., Kastner-Puschl S., Bolzano B., Kubista E. (1995), Land mobile Satellite narrowband propagation measurement campaign at Ka-Band, ESTEC contract 9949/92NL, Final report.
- ETSI (2008). *DVB-SH Implementation Guidelines*. TM-SSP252r9f.

- Fontan F., Vazquez-Castro M., Cabado C., Garcia J., Kubista E. (2001), Statistical modeling of the LMS channel, *IEEE Trans. on Vehicular Technology*, Vol. 50, No.6, Nov. 2001, 1549-1567, ISSN: 0018-9545.
- Loo C. (1985), A Statistical Model for a Land Mobile Satellite Link, *IEEE Trans. Vehicular. Technology*, Vol. VT-34, No.3, August 1985, 122-127, ISSN: 0018-9545.
- Fontan F., Vazquez-Castro M., Buonomo S., Baptista P., and Arbesser-Rastburg B. (1998), S-Band LMS propagation channel behavior for different environments, degrees of shadowing and elevation angles, *IEEE Trans. on Broadcasting*, Vol. 44, March 1998, 40-76, ISSN: 0018-9316.
- Loo C. (1991), Further results on the statistics of propagation data at L-band (1542 MHz) for mobile satellite communications, *Proc. of IEEE Vehicular Technology Conference*, pp. 51-56, ISSN: 1090-3038, May 1991, Saint Louis, USA.

Throughput Optimization for UWB-Based Ad-Hoc Networks

Chuanyun Zou

*School of Information Engineering, Southwest University of Science and Technology
China*

1. Introduction

The increasing demand for portable, high data-rate communications has stimulated search for new wireless technologies. Ultra-wideband impulse radio (UWB-IR) is an emerging radio technology that can support data rates of megabit-per-second, while maintaining low average-power consumption. UWB uses very short, carrier-less pulses of bandwidth on the order of a few Gigahertz. Over the past decade, many individuals and corporations began asking the United States Federal Communications Commission (FCC) for permission to operate unlicensed UWB systems concurrent with existing narrowband signals. In 2002, the FCC decided to change the rules to allow UWB system operation in a broad range of frequencies between 3.1 and 10.6 GHz. The FCC defines UWB as a signal with either a *fractional bandwidth* of 20% of the center frequency or 500 MHz (when the center frequency is above 6 GHz). The formula proposed by the FCC commission for calculating the fractional bandwidth is $2(f_H - f_L)/(f_H + f_L)$ where f_H represents the upper frequency of the -10 dB emission limit and f_L represents the lower frequency limit of the -10 dB emission limit. What makes UWB systems unique is their large instantaneous bandwidth and the potential for very simple implementations. Additionally, the wide bandwidth and potential for low-cost digital design enable a single system to operate in different modes as a communications device, radar, or locator. Taken together, these properties give UWB systems a clear technical advantage over other more conventional approaches in high multipath environments at low to medium data rates. Communication over UWB is particularly attractive due to its wide range of bit-rates, resilience to multi-path fading, accurate ranging ability, low transmission power requirements, and low probability of interception. After substantial progress in research on the UWB physical layer, in recent years, researchers began to consider the design of UWB networks [1]-[9]. The maximum allowable UWB transmission power is limited to a very small value, since UWB shares the same frequency band with other existing wireless communication systems. Consequently, short-distance communications are the main uses considered and UWB networks will likely often be ad hoc in nature. In an ad-hoc network each node has to have a routing function and it is essential to use multihop transmission to reach nodes further away. Since each node has a network control function, even if one of the nodes is not working properly, its influence on the whole network is quite limited. Therefore, ad-hoc networks are excellent with respect to robustness. Ad-hoc network do not require any infrastructure, a feature which allows for instant deployment and rerouting of traffic around failed or congested nodes. Since in ad-

hoc networks it is unnecessary to deploy base stations, the cost of an ad-hoc networks system is expected to be considerably lower than the corresponding cost of a cellular infrastructure. Furthermore, fault-tolerance (for example, due to richness of alternative routes ^[15]) of this type of networks is also significantly improved. Ad-hoc networks can be reconfigured to adapt its operation in diverse network environments. As the results of these characteristics, ad-hoc networks became of interest to the commercial and to the military markets. It is expected that UWB ad-hoc networks will be used for digital household electric appliances and peripheral equipment of PCs, for example, such as a wireless link between a PC and DVD player or a physical layer for a 'wireless USB' replacing traditional USB cables between devices. Examples of other applications that were considered are for networking among students in classrooms or among delegates at a convention centre. The mechanisms to best meet the requirements of the network layer for wireless ad hoc networks are a focus of current research and are certainly not well understood for UWB, which is a nascent networking technology. There are opportunities to leverage both radio link characteristics, using cross-layer design, and application requirements to optimize network layer protocols. For example, UWB devices in an ad hoc network may self-organize themselves into hierarchical clusters in ways that consider mutual interference, power conservation, and application connectivity requirements.

Throughput, which is defined as the bit rate of successfully received data, is a key performance measure for a data communication networks. In a wireless ad hoc network, throughput is a function of various factors, including the transmission power, the symbol rate (i.e., data rate), the modulation and the coding schemes, the network size, the antenna directionality, the noise and the interference characteristics, the routing and the multiple access control (MAC) schemes, and numerous other parameters. How to allocate resource and determine the optimal transmission power, transmission rate and schedule is a very challenging issue. There are several related papers ^{[2]-[8]} in the technical literature that study the throughput capacity and the optimization of UWB networks. They have suggested that: (1) an exclusion region around a destination should be established, where nodes inside the exclusion region do not transmit and the nodes outside the exclusion region can transmit in parallel ^[4], (2) the optimal size of the exclusion region depends only on the path-loss exponent, the background noise level, and the cross-correlations factor ^[6], (3) each node should either transmit with full power or not transmit at all ^[7], (4) the design of MAC is independent of the choice of a routing scheme ^[5].

In this chapter, we analyze and investigate the maximal total network throughput of UWB based ad hoc wireless networks. Understanding how this characteristic affects system performance and design is critical to making informed engineering design decisions regarding UWB implementation. The objectives of our work are: (1) to obtain theoretical results which demonstrate the dependencies among the maximum achievable throughput of a network, the number of active links in the network, the bit rate and the transmission power of active links, and other parameters, and (2) to determine the implications of these dependencies on the allocation and scheduling of the network resources. Our analysis show that the optimal allocation should: (1) allow the transmitters to either transmit at maximum power or be turned off, (2) allow more than one transmission when the maximum powers of the links are less than some value, which we term the critical power, (3) allow only one transmission when the maximum powers of the links are larger than the critical power, and (4) adjust the transmission rates to maintain the optimal transmission rates. We also derive an expression of the optimum transmission rate. As an example, we analytically calculate

the critical transmission power for the case of two-links and for the case of a scenario of N-links. Our results imply that the design of the optimal MAC scheme is not independent of the choice of the routing scheme. Furthermore, we expect our results obtain in this chapter to be helpful to network protocol design as well.

This chapter is organized as follows. The next section describes the UWB transmission system and formalizes the throughput optimization problem. In Section 3, we demonstrate the solution for the case of two simultaneous transmitters, while in Section 4 we analyze a network with arbitrary number of transmitters. Section 5 discusses the implications of the results, and the summary is given in Section 6.

2. Analytical model

We consider an ad hoc wireless network (Fig. 1.) that consists of identical nodes, each equipped with a half-duplex UWB radio. A transmitting node (a source node) is associated with a single receiver node (a destination node) and a pair of source-destination nodes forms a communication link. Each link can be selected for transmission by the MAC protocol based on some traffic requirements.

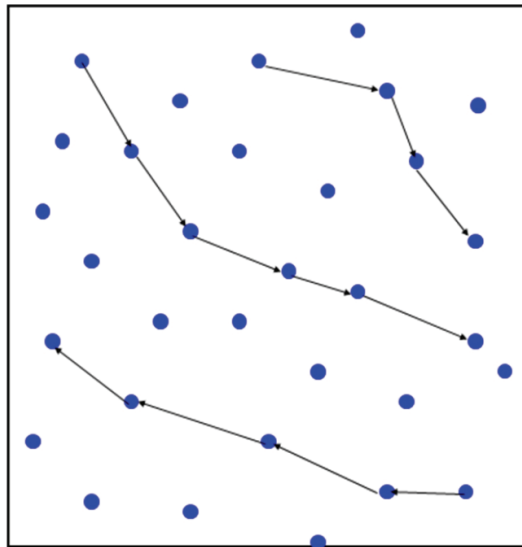


Fig. 1. A Multiple Hop Ad Hoc Network

We assume that the physical link layer is based on the *Time Hopping with Pulse Position Modulation* (TH-PPM) scheme, described in refs. [10--12]. In PPM, each monocycle pulse occupies a *frame*. Signal information is contained in pulse time position relative to the frame boundaries. Each bit is represented as L PPM-modulated pulses. An analytic TH-PPM representation of the transmitted signal of the k -th node is given by

$$s^k(t) = \sum_j w(t - jT_f - c_j^k T_c - \delta D_{\lfloor j/L \rfloor}^k) \quad (1)$$

where $w(t)$ denotes the monocycle pulse waveform, T_f is the nominal frame or pulse repetition interval, c^k_j is a user-unique pseudorandom TH code sequence (used for multiple access), T_c is the TH code chip period, $D^k_{\lfloor j/L \rfloor}$ is the k -th user's $\lfloor j/L \rfloor$ -th data symbol, where $\lfloor j/L \rfloor$ is the integer part of j/L and a symbol is transmitted as L monocycles PPM-modulated pulses, and δ is the amount of time shift of the PPM pulse for a data bit of "1".

The UWB communication system considered in this chapter is a spread-spectrum communication system, which uses a multiple-access scheme. Time hopping is used for multiple accesses. The source and the destination of each link have a common pseudorandom time hopping sequence, which is independent of other links' sequences. In the multiple-access scheme, transmissions on other links contribute added interference to the received signal and, due to randomness in time-hopping codes, we model such an interference as having statistical properties of Gaussian noise. The total noise at a receiver is comprised of background noise and a sum of interferences from all other active transmitters. The communication channel is assumed to be an AWGN channel. Thus, supposing that N links are active at a given time, the signal-to-interference plus noise (SINR) at the i -th link's receiver is represented as γ_i and is defined as [10]

$$\gamma_i = \frac{p_i g_{ii}}{R_i T_f \left(\eta_i + \rho \sum_{k=1, k \neq i}^N p_k g_{ki} \right)} \quad (2)$$

where R_i is the data transmission rate of i -th link and $R_i = 1/(LT_f)$, p_i is the average transmission power of the i -th link's transmitter, g_{ij} denotes path gain from the i -th link's transmitter to j -th link's receiver (g_{ii} is referred to as the i -th link's path gain and g_{ij} ($i \neq j$) is the interference path gain), η_i denotes the power of the background noise at i -th link's receiver, and ρ represents a parameter which depends on the shape of impulse((79) in ref. [10]).

In this work, a link is comprised of a pair of transmitter and receiver and the link is active if it is transmitting. When N links in a network are active at a given time, we define the throughput of the i -th link as the number of packets per second received without error at the i -th link's receiver:

$$T_i^N = R_i f(\gamma_i) \quad (3)$$

where $f(\gamma_i)$ is the packet success rate; i.e., it is the probability that the i -th link's receiver decodes a data packet correctly as a function of γ_i . The actual form of $f(\gamma_i)$ depends on the UWB receiver's configuration, the packet size, the channel coding, and the radio propagation model. We do not impose any restrictions on the form of $f(\gamma_i)$, except that $f(\gamma_i)$ is a smooth monotonically increasing function of γ_i , and $0 \leq f(\gamma_i) \leq 1$.

The total network throughput of N active links in the network, which we term T^N , is the sum of the N individual throughputs T_i^N .

$$T^N = \sum_{i=1}^N T_i^N \quad (4)$$

The aim of our optimization study is to determine the rate and the power assignments among the N links when the link gains and the background noise are given such that the total network throughput is maximized.

First we examine the properties of the throughput of link i , T_i^N , as a function of SINR. Using the following definition:

$$\mu_i = \frac{p_i g_{ii}}{T_f \left(\eta_i + \rho \sum_{k=1, k \neq i}^N p_k g_{ki} \right)} \quad (5)$$

eqs. (1) and (3) can now be represented respectively as

$$\gamma_i = \frac{\mu_i}{R_i} \quad (6)$$

$$T_i^N = \mu_i \frac{f(\gamma_i)}{\gamma_i} \quad (7)$$

Given the links' powers p_i ($i=1, \dots, N$), the value of μ_i is fixed and SINR γ_i varies only with rate R_i . As the rate R_i increases, the SINR γ_i and the packet success rate $f(\gamma_i)$ decrease. From eq. (7), we can see that too large or too small SINR leads to reduced throughput; at small SINR, the throughput is limited by small packet transmission success probability; however, at large SINR, the throughput is limited by small data transmission rate. Thus, we expect that there is an optimal value of SINR or an optimal symbol rate which corresponds to the maximum throughput.

3. Optimization for the two-links case

Before analyzing the performance of an arbitrary number of active links, we examine the case of two active links ($N=2$). This will allow us to gain some insight into the optimum allocation of transmission rates and transmission powers based on maximization of the throughput.

In the case of two active links, the total throughput is

$$T^2 = \mu_1 \frac{f(\gamma_1)}{\gamma_1} + \mu_2 \frac{f(\gamma_2)}{\gamma_2} \quad (8)$$

To obtain the optimal values of SINRs, γ_1^* and γ_2^* , that maximize the total network throughput, when p_1 and p_2 are fixed, we differentiate eq. (8) with respect to γ_1 and γ_2 , setting the first derivatives at zero and verifying that the second derivatives are negative. A simple calculation reveals that the conditions for both γ_1^* and γ_2^* are the same and, therefore, we can write $\gamma_1^* = \gamma_2^* = \gamma_c$ and state the conditions on γ_c as follows:

$$f(\gamma_c) = \gamma_c f'(\gamma_c) \quad (9)$$

$$f''(\gamma_c) < 0 \quad (10)$$

Then, from eq. (6), we calculate the optimal data rates:

$$\begin{aligned}
R_1^* &= \frac{\mu_1}{\gamma_c} = \frac{1}{\gamma_c T_f} \cdot \frac{g_{11} p_1}{\eta_1 + \rho g_{21} p_2} \\
R_2^* &= \frac{\mu_2}{\gamma_c} = \frac{1}{\gamma_c T_f} \cdot \frac{g_{22} p_2}{\eta_2 + \rho g_{12} p_1}
\end{aligned} \tag{11}$$

And with the above conditions, the optimal total network throughput is

$$\begin{aligned}
T^{2*} &= f'(\gamma_c)(\mu_1 + \mu_2) \\
&= \frac{f'(\gamma_c)}{T_f} \left(\frac{g_{11} p_1}{\eta_1 + \rho g_{21} p_2} + \frac{g_{22} p_2}{\eta_2 + \rho g_{12} p_1} \right)
\end{aligned} \tag{12}$$

When there is only a single active link in the network, either $p_2=0$ or $p_1=0$, the optimum total throughput is, respectively

$$\begin{aligned}
T_1^{1*} &= T^{2*}(p_2=0) = \frac{f'(\gamma_c)}{T_f} \cdot \frac{g_{11} p_1}{\eta_1} \\
T_2^{1*} &= T^{2*}(p_1=0) = \frac{f'(\gamma_c)}{T_f} \cdot \frac{g_{22} p_2}{\eta_2}
\end{aligned} \tag{13}$$

If we can adapt the transmission rates to the transmission powers according to eq. (11), the optimal total network throughput is then a function of the two links' powers and its value is determined by eqs. (12) and (13). Next, we show how to allocate the transmission powers between the two links so as to maximize the total network throughput. To do so, we focus our attention on eq. (12). From eq. (12), the optimal total network throughput is a function of p_2 only for fixed value of p_1 . In Figure 2, we depict a set of curves of the optimal total network throughput for different values of p_1 . Note that the graph includes the value of T_2^{1*} (i.e., $T^{2*}(p_1=0)$) and that the values for $p_2=0$ correspond to the situation in which only the first link is active. We state two observations: Firstly, we note that the throughput increases for large enough values of p_2 and that for small values of p_1 , the value of T^{2*} increases faster than for larger values of p_1 , so that T_2^{1*} will eventually exceed T^{2*} for non-zero p_1 . Secondly, we observe from the Figure that, there is a critical value, p_{c1} , such that if p_1 is larger than p_{c1} , T^{2*} will first decrease, take on a minimum, and then increase as p_2 grows. However, if p_1 is smaller than p_{c1} , T^{2*} will always be an increasing function of p_2 , with a minimum at $p_2=0$ (i.e., when the second link is inactive). These two observations imply that, when the two powers are high enough, the optimal total network throughput of two active links will always be smaller than the throughput of a single active link, but if the power of the first link is smaller than p_{c1} , then the adding of the second link increases the optimal total network throughput. To obtain the value of p_{c1} , we set $\partial T^{2*} / \partial p_2$ at $p_2=0$ at zero, which results in

$$p_{c1} = \frac{\eta_2}{2\rho g_{12}} \left(\sqrt{1 + 4 \frac{\eta_1^2 g_{12} g_{22}}{\eta_2^2 g_{21} g_{11}}} - 1 \right) \tag{14}$$

Also, if eq. (12) is seen as a function of single variable p_1 with p_2 being a parameter, we can obtain the critical value of p_2 as

$$p_{c2} = \frac{\eta_1}{2\rho g_{21}} \left(\sqrt{1 + 4 \frac{\eta_2^2 g_{21} g_{11}}{\eta_1^2 g_{12} g_{22}}} - 1 \right) \tag{15}$$

When p_2 is smaller than p_{c2} , T^* will always be an increasing function of p_1 . If the power p_1 and p_2 simultaneously satisfy the following two inequalities: $p_1 < p_{c1}$ and $p_2 < p_{c2}$, then the total network throughput, T^* , is larger than the throughputs of the single active link case with the same power, T_{11}^* and T_{21}^* . In the example of Figure 2, we find that $p_{c1}=90.95$ mW and $p_{c2}=155.69$ mW.

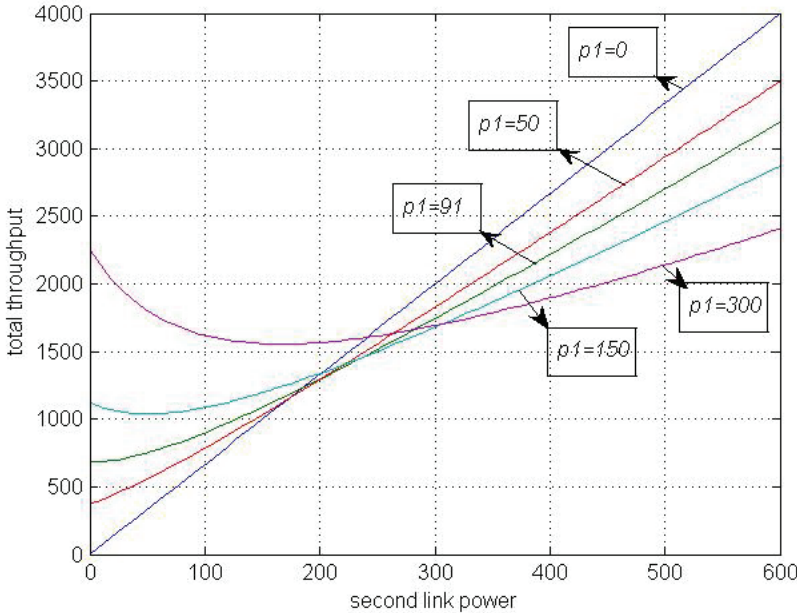


Fig. 2. The maximal total throughput vs. the power of the second link, the power of the first link as the parameter, and with the following values of parameters in (12): $f(\gamma_c)/T_f=1$ bit/s, $g_{11}=0.03$, $g_{22}=0.04$, $g_{21}=0.003$, $g_{12}=0.002$, $\rho=0.01$, $\eta_1=0.004$ mW, $\eta_2=0.006$ mW, $p_{c1}=90.95$ mW, $p_{c2}=155.69$ mW

In any practical situation, transmission powers are not unlimited. But, using eq. (11), we can calculate the corresponding optimal transmission rates according to the attainable transmission power values and, so as to achieve the optimal throughput. We describe how to allocate the transmission powers, so as to maximize the throughput, when $0 < p_1 < P_1$ and $0 < p_2 < P_2$. Since the sign of the second derivatives of eq. (12) with respect to p_1 and p_2 is positive for any value of p_1 and p_2 , the maximum throughput lies on the boundary of the attainable region, i.e., $[0 < p_1 < P_1, 0 < p_2 < P_2]$. Based on our analytic results obtained so far, if $P_1 < p_{c1}$ and $P_2 < p_{c2}$, the optimum transmission power allocation is $p_1=P_1$ and $p_2=P_2$, i.e., the two links' transmitters transmit at their maximum powers and at the same time (Figure 3 is an example of such a case). However, if $P_1 > p_{c1}$ and $P_2 > p_{c2}$, the optimum allocation is $p_1=P_1$, $p_2=0$ or $p_1=0$, $p_2= P_2$, i.e., the transmitter of one link transmits at its maximum power, while the other is turned off (Figure 4 is an example of such a case).

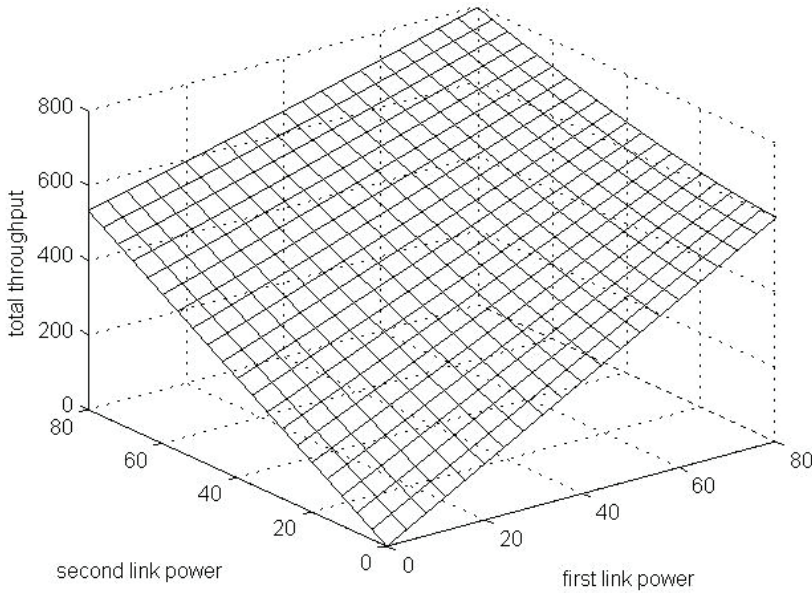


Fig. 3. The maximal total throughput vs. the transmission powers of the two link, when maximum attainable powers are smaller than the critical values, for the same parameters' values as in Figure 2 (the total throughput is maximum at $p_1=80$ mW, $p_2=80$ mW)

A transmitted signal attenuates according to a power law as a function of distance from its transmitter; i.e., if d_{ij} is the distance from the i -th link's transmitter to j -th link's receiver, then

$$g_{ij} = c \cdot d_{ij}^{-\alpha} \tag{16}$$

where c and a are constants. This is a commonly used attenuation model for wireless transmissions, and it has been verified as applicable to an UWB indoor propagation model [13][14]. Hence, p_{c1} and p_{c2} are functions of d_{12} , d_{21} , d_{11} , and d_{22} . From eqs. (14) and (15), we calculate the two critical distances, d_{c12} and d_{c21} for given values of P_1 , P_2 , d_{11} , d_{22} , and either d_{12} or d_{21} .

$$d_{c21} = \frac{d_{22}}{d_{11}} \cdot \left[\frac{(\rho c d_{12}^{-\alpha} P_1^2 + \eta_2 P_1) \rho c}{\eta_1^2} \right]^{\frac{1}{\alpha}} \tag{17}$$

$$d_{c12} = \frac{d_{11}}{d_{22}} \cdot \left[\frac{(\rho c d_{21}^{-\alpha} P_2^2 + \eta_1 P_2) \rho c}{\eta_2^2} \right]^{\frac{1}{\alpha}} \tag{18}$$

So, if $d_{12} < d_{c12}$ or $d_{21} < d_{c21}$, only one link should be active. This conclusion is equivalent to the concept of "the exclusion regions" in refs. [4--6], but in our case the exclusion regions sizes, d_{c12} and d_{c21} , depend on the transmission powers of the sources, the powers of background noises, the path-loss exponent, and the length of the links; thus our solution is different from the proposition in refs. [4--6].

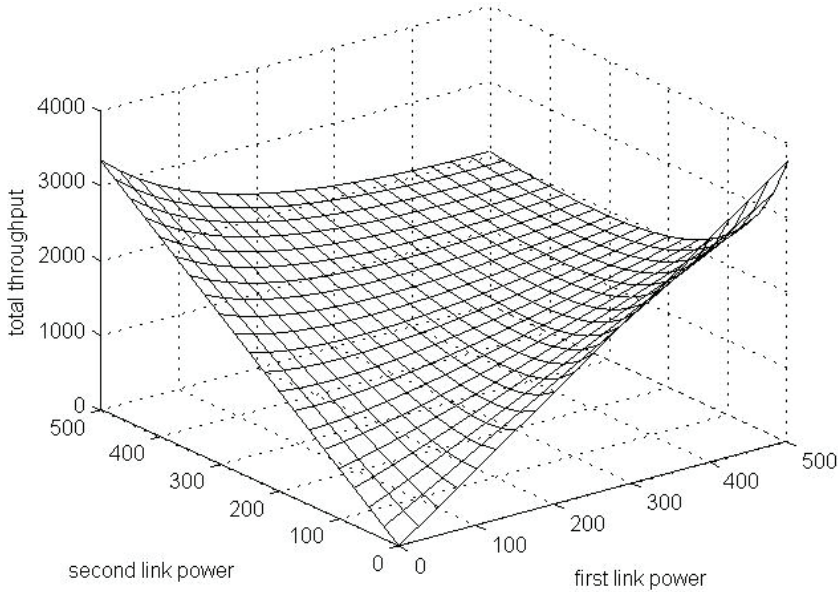


Fig. 4. The maximal total throughput vs. the transmission powers of the two link, when maximum attainable powers are larger than the critical values, for the same parameters' values as in Figure 2 (the total throughput is maximum at $p_1=500$ mW, $p_2=0$ mW)

4. Optimization for N links

We now expand our study to consider the optimization problem of eq. (4) for networks with N active links. Examining the first and second derivatives of (4) with respect to γ_i ($i=1, \dots, N$), we find that all the optimal values of SINRs, γ_i^* ($i=1, \dots, N$), correspond to one and the same value, γ_c , a value which satisfies eq. (9) and (10). So the N optimal rates are

$$R_i^* = \frac{\mu_i}{\gamma_c} = \frac{1}{\gamma_c T_f} \cdot \frac{g_{ii} p_i}{\eta_i + \rho \sum_{k=1, k \neq i}^N g_{ki} p_k} \quad i = 1 \dots N \quad (19)$$

Accordingly, the optimum total network throughput is

$$T^{N*} = \frac{f'(\gamma_c)}{T_f} \sum_{i=1}^N \frac{g_{ii} p_i}{\eta_i + \rho \sum_{k=1, k \neq i}^N g_{ki} p_k} \quad (20)$$

We fix all p_i ($i=1, \dots, N$) at some arbitrary values, except for p_j , and we consider eq. (20) as a function of a single free variable p_j . We can draw curves similar to those in Figure 2, but the values for $p_j=0$ are now the throughputs of the $N-1$ active links. The first and second partial derivatives of (20) with respect to p_j are

$$\frac{\partial T^{N*}}{\partial p_j} = \frac{f'(\gamma_c)}{T_f} \left[\frac{g_{jj}}{\eta_j + \rho \sum_{k=1, k \neq j}^N g_{kj} p_k} - \sum_{i=1, i \neq j}^N \frac{\rho g_{ji} g_{ii} p_i}{\left(\eta_i + \rho \sum_{k=1, k \neq i}^N g_{ki} p_k \right)^2} \right] \quad j=1 \dots N \quad (21)$$

$$\frac{\partial^2 T^{N*}}{\partial p_j^2} = \frac{f'(\gamma_c)}{T_f} \sum_{i=1, i \neq j}^N \frac{2\rho g_{ji}^2 g_{ii} p_i}{\left(\eta_i + \rho \sum_{k=1, k \neq i}^N g_{ki} p_k \right)^3} \quad j=1 \dots N \quad (22)$$

Because eq. (22) is always positive for any $p_j > 0$ ($j=1 \dots N$), T^{N*} is always a concave function, and hence its maximum is only attained either at $p_j=0$ or at the value of maximum transmission power, $p_j=P_j$. Of course, $p_j=0$ means that the link j is inactive, while $p_j=P_j$ means transmission at maximum attainable power. So to solve the maximal throughput problem, we need to determine how many links will be active (transmitting with maximal power). By setting eq. (21) at $p_j=0$ ($j=1 \dots N$) at zero, we can compute a set of critical p_{cj} ($j=1 \dots N$). When $P_j < p_{cj}$ ($j=1 \dots N$), since eq. (21) is always positive, then the maximal total throughput of N active links, T^{N*} , is larger than the maximal throughput of single active link, T^1 , and larger than the maximal throughput of $N-1$ active links, $T^{(N-1)*}$. So the optimal scheduling is to allow all the N links to transmit, each at its maximal power. When $P_j > p_{cj}$ ($j=1 \dots N$), the maximal total throughput of N active links might be less than the maximal throughput of a single active link. So, at any particular time, the optimal scheduling should allocate transmission of one active link with large enough power, while the other transmitters are turned off. We could also arrive at this conclusion by the following argument. If we allocate each link's transmitting power as $p_i = a_i p$ ($i=1 \dots N$), a_i being a positive constant or zero, then eq. (20) becomes

$$T^{N*} = \frac{f'(\gamma_c)}{T_f} \sum_{i=1}^N \frac{g_{ii} a_i}{\frac{\eta_i}{p} + \rho \sum_{k=1, k \neq i}^N g_{ki} a_k} \quad (23)$$

We can see that T^{N*} is an increasing function of p . When p is large enough (strictly, infinity), we can obtain

$$T^{N*} = \frac{f'(\gamma_c)}{T_f} \sum_{i=1}^N \frac{g_{ii} a_i}{\rho \sum_{k=1, k \neq i}^N g_{ki} a_k} \quad (24)$$

When more than two links are active, the value of T^{N*} is limited. However, if just one link is active, for example, $a_i=0$ ($i=2 \dots N$) but $a_1 \neq 0$, then T^{N*} tends to infinity.

We consider a special scenario when $g_{ii}=g$, $g_{ij}=g' (i \neq j)$, $\eta_i=\eta$, and $p_i=p$ ($i,j=1 \dots N$). With these conditions, the single active link's maximal throughput is

$$T^{1*} = \frac{f'(\gamma_c) g p}{T_f \eta} \quad (25)$$

However, the maximal total network throughput of N active links is in this case:

$$T^{N*} = \frac{f'(\gamma_c)}{T_f} \frac{Ng}{\frac{\eta}{p} + (N-1)\rho g'} \quad (26)$$

and each link's maximal throughput is

$$T_1^{N*} = \frac{f'(\gamma_c)}{T_f} \frac{g}{\frac{\eta}{p} + (N-1)\rho g'} \quad (27)$$

If we let p go to infinity, T^{1*} will approach infinity as well, but T^{N*} approaches the following finite value:

$$T^{N*} = \frac{N}{N-1} \cdot \frac{f'(\gamma_c)}{T_f} \cdot \frac{g}{\rho g'} \quad (28)$$

We can also see that T^{N*} is an increasing function of N . So with N increasing to infinity, eq. (27) decreases to zero, but eqs. (26) and (28) approach

$$T^{\infty*} = \frac{f'(\gamma_c)}{T_f} \cdot \frac{g}{\rho g'} \quad (29)$$

From comparison, eq. (26) will be smaller than eq. (25) when p is larger than the following value of p_c :

$$p_c = \frac{\eta}{\rho g'} \quad (30)$$

Using eq. (16), we calculate the critical value of the interference distance, d'_c , for transmitted power p :

$$d'_c = \left(\frac{\rho c p}{\eta} \right)^{\frac{1}{\alpha}} \quad (31)$$

In this special symmetric scenario, the critical power, p_c , is independent of N . and the critical interference distance, d'_c , is independent of the link length. When $0 < p < p_c$, the maximal total network throughput is larger than the maximal single active throughput and the increment, $T^{N*} - T^{1*}$, is maximum when p is equal to the following value of p_m :

$$p_m = \frac{\eta}{(\sqrt{N} + 1)\rho g'} \quad (32)$$

When $p=p_m$, eqs. (25) and (26) become, respectively

$$T_m^{1*} = \frac{f'(\gamma_c)}{T_f} \cdot \frac{g}{(\sqrt{N}+1)\rho g'} \quad (33)$$

$$\begin{aligned} T_m^{N*} &= \frac{\sqrt{N}}{\sqrt{N}+1} \cdot \frac{f'(\gamma_c)}{T_f} \cdot \frac{g}{\rho g'} \\ &= \frac{\sqrt{N}}{\sqrt{N}+1} T^{\infty*} \\ &= \sqrt{N} T_m^{1*} \end{aligned} \quad (34)$$

and the maximal throughput of each link is

$$\begin{aligned} T_1^{N*} &= \frac{T_m^{N*}}{N} = \frac{T^{\infty*}}{N + \sqrt{N}} \\ &= \frac{T_m^{1*}}{\sqrt{N}} \end{aligned} \quad (35)$$

Actually, $T^{\infty*}$ is the maximal total network throughput capacity of a network with concurrently active links, and 90% of the maximal total network throughput can be attained when $N=81$. From eq. (29), the maximal total network throughput $T^{\infty*}$ is mainly determined by the physical layer, and it can be enhanced by increasing g (the signal gain) and $f'(\gamma_c)$ (packet transmission success probability increment rate at optimal SINR), and by decreasing T_f (pulse repetition interval), ρ (the shape factor of impulse), and g' (the interference gain). The values of $f'(\gamma_c)$, T_f , and ρ depend on design parameters, such as modulation, pulse shape, time-hopping sequences, and the size of data packets. The values of g and g' depend on the antenna design; e.g., multiple transmit and receive antennas (MIMO) [16] can increase g and decrease g' . However, g and g' are also affected by the routing and the MAC schemes.

5. Discussion and concluding remarks

While the fundamental principles of networking are the same regardless of the underlying physical layer, UWB has unique characteristics that influence how protocols and a UWB system are designed. A UWB network can be represented by a five-layer model, compatible with the TCP/IP suite, that includes a UWB physical layer, associated data link layer, network layer, transport layer, and application layer. Each layer provides services to the layer directly above it and uses services provided by the layer beneath it. The unique characteristics of the UWB physical layer have the greatest influence on the design of the data link layer. The characteristics of the physical layer and the design of the associated data link layer may also influence the design of the network layer, transport layer, and even application layer, especially if a design is to achieve optimal performance.

When designing a communication network, it is important to understand how much information such a network can transport, what parameters affect the maximal throughput of the network, and how to change the parameters so as to maximize the throughput. The two last sections provide us with some answers to these questions for UWB wireless ad hoc

networks. We have established the dependencies among the maximum achievable throughput of the network, each active link's transmission rate and transmission power, the number of simultaneously active links in the network, the link and the interference paths gains, and the background noise.

In the MAC protocol of data link layer, time is divided into time slots, which are allocated for links according to the link-scheduling policy. There are two types of link-scheduling policies: single link policy which allows only one link to transmit in any slot, and concurrent links policy which allows multiple links to transmit simultaneously in a slot. These two policies require that the transmission rate of the active links be maintained at the optimal value according to eq. (19). Under this condition, the maximum total network throughput depends on each link's maximum power and on the interferences among the active links. Our results show that the single link policy suits transmissions with large power: the throughput increases linearly with the power (and, in theory, indefinitely), as shown in eq. (13) and (25). With this policy, the larger is the power, the larger is the throughput. With the concurrent links policy, the maximal total network throughput cannot increase indefinitely by continual increase in transmission powers. Actually, the maximal total network throughput is limited by the interference levels among the active links, and the throughput approaches a finite value when multiple powers are increased indefinitely. This is demonstrated by eqs. (24) and (28). So, on one hand, when the powers are large enough, the single active link maximal throughput exceeds the maximal total network throughput of concurrently active links. In this situation, it is better to choose the single link policy. On the other hand, the maximal total network throughput of the concurrently active links is larger than the maximal throughput of a single active link, if each link power is below the critical value or when the separation between any pair of active links is above their critical values. These critical values are computed in eqs. (14), (15), (30), (17), (18), and (31). Hence, the concurrent links policy is suitable for small powers or for sparse networks. In this situation, each link has an optimum power value which maximizes the throughput gain by increasing the number of the concurrently active links.

The maximal total network throughput with concurrent active links, or the network capacity, is calculated by eq. (29), and can be enhanced by decreasing the interference path gains or by increasing the link path gain, but not by increasing the power. As the number of concurrently active links, N , is increasing, each link throughput is decreased. Existing protocols (like 802.11) are based on the single link policy, but their rate might not be optimum. The regulatory bodies (like FCC) impose severe limitation on UWB power density to avoid interference on other existing wireless communication systems (such as GPS and 802.11 networks), since they share the same frequency band. The FCC regulation allows commercial UWB devices to emit no more than -41 dBm/MHz of average transmitted power, so the maximum transmitted power is limited to less than -2.2 dBm, or approximately half a Milliwatt. Consequently, the concurrent links policy may be a more suitable choice for UWB ad hoc networks.

Because the routing protocol of network layer determines the paths of data flow and interference gains between intended links, the design of an MAC protocol based on the concurrent links policy should be related to the choice of a routing protocol for maximization of the total network throughput. However, the design of an MAC protocol based on the single link policy should be independent of the choice of routing protocol. As the rate adaptation requires support of the physical layer, such adaptation is most efficiently

performed if the design is based on cross-layer considerations. The application of our results to implementation of an MAC protocol based on the concurrent links policy with cross-layer design considerations is outside the scope of this chapter, but is left for future study.

When a mechanism for adaptation of transmission rates is incorporated into the design of the MAC protocol, by adjusting the transmission rates to their optimum values, the maximal total network throughput is limited by maximal transmission power and by the interference from other active links in the networks. The maximal total network throughput approaches a constant and each link's throughput approaches zero as the maximal transmission power and the simultaneously active links increase in number. For the case of a single active link, the maximal throughput increases linearly with the maximal transmission power and, barring a limit on transmission power, the maximal throughput can increase indefinitely. When the values of the maximal transmission power are large enough, the maximal throughput in the single active link case exceeds the maximal total network throughput of the multiple active links case. To maximize the total network throughput, the optimal transmission scheduling should allocate at any time transmission on one link only when the maximal transmission power is large and the interference is strong. However, when the maximal transmission power is small and the interference is weak, the optimal transmission scheduling should allocate at any time simultaneous transmission on multiple links.

6. Summary

In this chapter, we study the problem of radio resource allocation, both transmission rates and transmission powers, so as to maximize the throughput of UWB wireless ad-hoc networks. Our analysis is based on the packet-success function (PSF), which is defined as the probability of a data packet being successfully received as a function of the receiver's signal-to-interference-and-noise-ratio (SINR). We find an optimal link transmission rate, which maximizes the link's throughput and is dependent on the all active links transmission powers. If each link transmission rate is adapted to this optimal link transmission rate, then, with single-link operation (i.e., no other interference sources are present), the link's throughput is directly proportional to the transmitter's power and increases indefinitely with increasing transmission power. However, with multiple-links operation and interference each other, as each link transmitting power increases, so does the interference level, and the total network throughput approaches a constant other than infinite. Thus, for sufficiently small transmission power, the total network throughput of the multiple-links case exceeds the throughput of the single-link case, but the reverse happens for high power. In addition, this chapter reveals that, as the number of concurrently transmitting links increases, regardless of the power level, the maximal total network throughput approaches a constant, with each link's throughput approaching zero. To maximize the network throughput, for the case of small maximal transmission power with weak interference levels, the optimal transmission scheduling allocates simultaneous transmissions of multiple links, but for the case of large maximal transmission power with strong interference levels, the optimal policy assigns separate time for transmission on each link. The breakpoint of when to use one link or multiple links is termed the critical power. As an example of the analytical calculation of the critical link's power, we present here solutions for a two-link case and an N -link case. In contrast with previous studies, our results imply that the design of optimal MAC is dependent on the choice of a routing scheme.

7. Acknowledgement

This work is supported by the Scientific Foundation of Sichuan Education Department of China (Grant No. 2006A096), the Ph. D Foundation of Southwest University of Science and Technology (Grant No. 06zx7107), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Personnel Ministry (Grant No. 08ZD0106)

8. References

- [1] Apsel A, Dokania R, Wang X. Ultra-low power radios for ad-hoc networks. In: IEEE International Symposium on Circuits and Systems, Taipei, 2009. 1433 - 1436
- [2] Tang X, Hua Y. Capacity of ultra-wideband power-constrained Ad Hoc networks. *IEEE Trans Inf Theory*, 2008, 54(2): 916 - 920
- [3] Zou C, Haas Z. Optimal Resource Allocation for UWB Wireless AD HOC Networks. In: IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, Berlin, 2005. 452-456
- [4] Radunovic B, Boudec J-Y L. Optimal power control, scheduling, and routing in UWB networks. *IEEE J Sel Area Commun*, 2004, 22(7): 1252-1270
- [5] Merz R, Boudec J-Y L, Widmer J, et al. A rate-adaptive MAC protocol for low-power ultra-wide band ad-hoc networks. In: Proceeding of 3rd International Conference on AD-HOC Networks and Wireless, Vancouver, 2004. 306-311
- [6] Liu K H, Cai L, Shen X. Exclusive-region based scheduling algorithms for UWB WPAN. *IEEE Trans Wireless Commun*, 2008, 7(3): 933 - 942
- [7] Cuomo F, Martello C, Baiocchi A, et al. Radio resource sharing for ad hoc networking with UWB. *IEEE J Sel Area Commun*, 2002, 20(12): 1722-1732
- [8] Baldi P, Nardis L D, Benedetto M-G D. Modeling and optimization of UWB communication networks through a flexible cost function. *IEEE J Sel Area Commun*, 2002, 20(12): 1733-1744
- [9] Zhu S, Leung KK, Constantinides A G. Distributed cooperative data relaying for diversity in impulse-based UWB Ad-Hoc networks. *IEEE Trans Wireless Commun*, 2009, 8(8): 4037 - 4047
- [10] Win M, Scholtz R. Ultra-wide bandwidth time-hopping spread spectrum impulse radio for wireless multiple-access communications. *IEEE Trans Commun*, 2000, 48(4): 679-691
- [11] Feng D, Ghauri S, Zhu Q. Application of the MUI model based on packets collision (PC) in UWB ad-hoc network. In: International Conference on Networking, Sensing and Control, 2009. 554 - 558
- [12] Durisi G, Romano G. On the validity of Gaussian approximation to characterize the multiuser capacity of UWB TH PPM. In: Proceeding of IEEE Conference of Ultra-Wideband Systems Technologies, 2002. 151-161
- [13] Molisch A F. Ultrawideband propagation channels-theory, measurement, and modeling. *IEEE Trans Veh Technol*, 2005, 54(5): 1528-1545
- [14] Ghassemzadeh S S, Tarokh V. UWB path loss characterization in residential environments. In: Proceeding of IEEE Radio Frequency Integrated Circuits (RFIC) Symposium, 2003: 501-504

- [15] Tsirigos A, Haas Z J. Analysis of multipath routing - Part I: The effect on the packet delivery ratio. *IEEE Trans Wireless Commun*, 2004, 3(1): 138-146, January
- [16] Kaiser T, Zheng F, Dimitrov E. An overview of ultra-wide-band systems with MIMO. *Proceedings of the IEEE*, 2009, 97(2): 285-312

Outage Probability Analysis of Cooperative Communications over Asymmetric Fading Channel

Sudhan Majhi, Youssef Nasser and Jean François H elard
*National Institute of Applied Sciences of Rennes
France*

1. Introduction

Cooperative relaying is a promising technology for future wireless communications. It is mostly applicable to the small dimensional and limited power devices, which cannot use the conventional multiple input multiple output (MIMO) technology to obtain the advantages of MIMO. It can benefit most of the leverages of MIMO such as array gain, diversity gain, spatial multiplexing gain and interference reduction without using the conventional MIMO technology Liu et al. (2009); Laneman et al. (2004); Paulraj et al. (2004). Since the original signal is forwarded by relay nodes, the performance of the relaying network depends on the relaying process of the relay nodes and fading characteristic of their links. Classically, relay network has three links source-destination (S-D), source-relay (S-R) and relay-destination (R-D) and the relaying processes are classified as amplify-and-forward (AF), decode-and-forward (DF) and compress-and-forward (CF) Krikidis & Thompson (2008); Nosratinia et al. (2004).

The diversity of the relaying network depends on the degree of freedom of the network. For repetition-based relaying, the degree of freedom increases with the number of relay nodes when the system is an half duplex and use a time division duplex Zhao et al. (2005, 2007). However, it suffers spectral efficiency with increase in the number of relay nodes in the network. On the other hand, opportunistic relaying uses only one relay node to forward source data to the destination and its degree of freedom is two. It has higher spectral efficiency and better outage performance than that of repetition-based relaying.

The performance of the relaying network depends on the fading characteristic of S-D, S-R and R-D links, i.e. diversity of the relaying networks depends on the location of the relay nodes and its surrounding environment. In practice, cooperative nodes are usually located in different geographical locations and at different distances with respect to S and D. The signal in one link may be in line of sight (LOS) situation and other links may be in NLOS situation. For example, fixed relay nodes are used for forwarding source's data to a specific region (e.g. tunnel, behind of the building) and they often use directional antennas, so the R-D link is likely to be a LOS situation. However, we cannot assume such scenario for other links specially when D is in a shadowing region with respect to S. In other words, one link may undergo Rician fading channel and others links may undergo Rayleigh fading channel. Such scenario is refereed to as asymmetric fading channel. This channel scenario can also be

seen in cooperative cognitive radio where secondary terminal works as a relay. In addition, all the links, i.e., S to i^{th} relay (S- R_i) and i^{th} relay to D (R_i -D), may be independent but non-identically distributed (i.n.d) fading channels. Therefore, a complete outage performance study of cooperative relaying for such asymmetric and i.n.d fading channels is required.

The outage probability of relaying networks over symmetric fading channel, in which all the links undergo the same fading distribution, is provided in several works Hwang et al. (2007); Xu et al. (2009); Zhao et al. (2006); Savazzi & Spagnolini (2008); Michalopoulos & Karagiannidis (2008); Zou et al. (2009); Vicario et al. (2009). The outage probability for repetition-based AF relaying over Rayleigh channel is provided in Zhao et al. (2007). The outage probability of AF relaying over Rician fading is provided only in few articles Zhu et al. (2008). The asymmetric fading channel, mix of Rayleigh and additive white Gaussian noise, is introduced in Katz & Shamai (2009). The performance of AF relaying over asymmetric fading, mix of Rician & Rayleigh, is provided in independent work in Suraweera et al. (2009); Suraweera, Karagiannidis & Smith (2009). However, in the literature, none of the papers provided any closed form outage probability of AF relaying over asymmetric fading channels.

In this work, we provide a complete study of outage probability of repetition-based and opportunistic relaying over asymmetric fading channels. The closed form of outage probability is derived over i.n.d fading channel at high SNR regime. In this work, we adopted AF relaying networks over two different scenarios, called asymmetric channel I and asymmetric channel II given in Fig. 1. We provide analytical model of each of the asymmetric channel and verified through the Monte-Carlo simulation studies. The obtained results of asymmetric fading channel are compared with symmetric fading channel, i.e., with Rician fading channel and Rayleigh fading channel. When outage performance is compared between two asymmetric channels, asymmetric channel I provides better outage performance than asymmetric channel II and when it is compared between two diversity techniques, opportunistic AF relaying provides better outage performance than the repetition-based AF relaying.

The rest of the article is organized as follows. Section 2 discusses a two-hop AF relaying network and asymmetric channel models. Section 3 derives the outage probability of the repetition-based AF relaying over two different asymmetric fading channel scenarios at high SNR regime. Section 4 derives the outage probability of opportunistic AF relaying over asymmetric channel I and asymmetric channel II. Finally, conclusion is drawn in section 6.

2. System model

2.1 Signal model of AF relaying

In this framework, we consider a general 2-hop AF relaying network consisting of S, M relays, R_i , $i=1, 2, \dots, M$, and D. We assume that D performs maximal ratio combining at the receiver. The network has $M+1$ time slots for M relay nodes Zhao et al. (2005). In the first time slot, S broadcasts data to D and all R_i . The received signals at D and R_i are given by

$$y_{sd} = h_{sd}x + \eta_d \quad (1)$$

$$y_{sr_i} = \alpha h_{sr_i}x + \eta_{r_i} \quad (2)$$

where x is the signal transmitted by S, η_d and η_{r_i} are the zeromean complex Gaussian random variables at D and i^{th} relay, respectively. h_{sd} and h_{sr_i} are the fading coefficients of S-D and S- R_i links, respectively.

The received signal at D from R_i at the $(i + 1)^{\text{th}}$ time slot is

$$y_{r,d} = h_{r_i,d}x' + \eta_d \tag{3}$$

where $\alpha = 1$, x' is the signal transmitted by the relay R_i for the case of repetition-based relaying and $h_{r_i,d}$ is the fading coefficient of R_i -D link. For the opportunistic AF relaying, α is the amplifying factor and $x' = x$. In the 2^{nd} time slot, the best opportunistic relay node forwards the source's signal to D.

2.2 Channel model of asymmetric fading channel

For simplicity, we use different notations of the random variables for different fading distributions. For the Rayleigh channel, in general, let $\gamma_{ab} = P_s |h_{ab}|^2$ be the instantaneous signal power of a-b link and for the Rician fading channel, the corresponding instantaneous signal power is denoted as ξ_{ab} . The transmitted power from source and relay is P_s . The probability density function (PDF) of γ_{ab} and ξ_{ab} are expressed respectively as

$$f_{\gamma_{ab}}(x) = \frac{1}{\bar{\gamma}_{ab}} e^{-x/\bar{\gamma}_{ab}} \tag{4}$$

$$f_{\xi_{ab}}(\xi) = \frac{K_{ab} + 1}{\bar{\xi}_{ab}} e^{-\xi(K_{ab} + 1)/\bar{\xi}_{ab} - K_{ab}} I_0 \left(\sqrt{\frac{4K_{ab}(K_{ab} + 1)\xi}{\bar{\xi}_{ab}}} \right) \tag{5}$$

where $I_0(\cdot)$ is the 0^{th} order modified Bessel function of first kind, $\bar{\gamma}_{ab} = E\{\gamma_{ab}\}$, $\bar{\xi}_{ab} = E\{\xi_{ab}\}$, and K_{ab} is the Rician factor.

Although there are several possibilities of asymmetric fading channel, in this work, we assume two asymmetric fading channels: namely asymmetric channel I and asymmetric channel II, shown in Fig. 1. For the asymmetric channel I, we assume that S-R link undergoes Rayleigh distribution and S-D and R-D links undergo Rician fading distribution. For the asymmetric channel II, S-R link undergoes Rician distribution and S-D and R-D links undergo Rayleigh distribution.

3. Repetition based AF relaying

The repetition-based AF relaying is introduced in Laneman et al. (2004). Due to the higher degree of freedom of repetition-based relaying and simple implementation for AF relaying, it gain its own importance in cooperative communications. The equivalent instantaneous end-to-end signal-to-noise ratio (SNR) for repetition-based AF relaying is given as Zhao et al. (2006)

$$\gamma = \frac{P_s |h_{sd}|^2}{N_{sd}} + \sum_{i=1}^M \frac{\frac{P_s |h_{sr_i}|^2}{N_{sr_i}} \frac{P_s |h_{r_i,d}|^2}{N_{r_i,d}}}{\frac{P_s |h_{sr_i}|^2}{N_{sr_i}} + \frac{P_s |h_{r_i,d}|^2}{N_{r_i,d}} + 1} \tag{6}$$

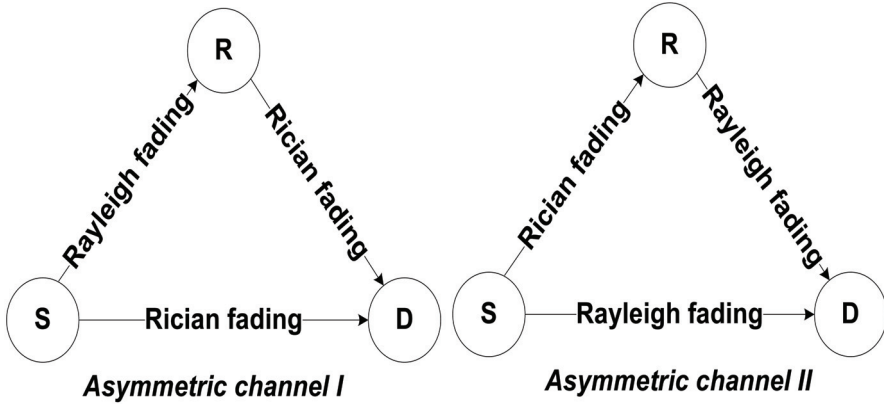


Fig. 1. Different asymmetric fading channels of a cooperative networks
 The upper bound of instantaneous SNR for the above can be written as

$$\gamma_{max} = P_s |h_{sd}|^2 \gamma_0 + \sum_{i=1}^M \min(P_s |h_{sr_i}|^2 \gamma_0, P_s |h_{r_i d}|^2 \gamma_0) \tag{7}$$

In the following section, we provide the lower bound of outage probability of the opportunistic AF relaying for different asymmetric fading channels shown in Fig. 1.

3.1 Asymmetric channel I:

For the asymmetric channel I, S-R link experiences Rayleigh distribution and S-D and R-D links experience Rician fading distribution, so (7) can be written as

$$\gamma_{max} = \gamma_0 \gamma_{sd} + \gamma_0 \xi_{sum} \tag{8}$$

where $\gamma_{sd} = P_a |h_{sd}|^2$ is the exponential distribution, $\xi_{sum} = \sum_{i=1}^M \xi_{min,i}$, $\xi_{min,i} = \min(\xi_{sr_i}, \xi_{r_i d})$, ξ_{sr_i} and $\xi_{r_i d}$ are the random variables of noncentral Chi-square distribution. The corresponding outage probability can be defined as

$$p_{out} = Pr[\gamma_{ub} < \gamma] \tag{9}$$

where $\gamma_{ub} = \gamma_{max} / \gamma_0$, $\gamma = (2^{(M+1)R} - 1) / \gamma_0$. The outage probability provided in (9) is equivalent to the commutative distribution function (CDF) of γ_{ub} . The direct evaluation CDF of γ_{ub} is complicated, so we use the initial value theorem (IVT) of Laplace transformation (LT). Therefore, this derived analytical results are valid for high SNR regime. To evaluate this, first CDF of $\xi_{min,i}$ needs to be evaluated. The CDF of the random variable $\xi_{min,i}$ can be written as

$$\begin{aligned} F_{\xi_{min,i}}(\gamma) &= 1 - (1 - Pr[\xi_{sr_i} < \gamma]) (1 - Pr[\xi_{r_i d} < \gamma]) \\ &= 1 - Q_1 \left(\sqrt{2K_{sr_i}}, \sqrt{\frac{2(K_{sr_i} + 1)\gamma}{\xi_{sr_i}}} \right) Q_1 \left(\sqrt{2K_{r_i d}}, \sqrt{\frac{2(K_{r_i d} + 1)\gamma}{\xi_{r_i d}}} \right) \end{aligned} \tag{10}$$

where $Q_1(\cdot)$ is the Marcum Q-function of first order. The PDF of $\xi_{min,i}$ is obtained by differentiating the above, which can be expressed as

$$f_{\xi_{min,i}}(\gamma) = Q_1\left(\sqrt{2K_{sr_i}}, \sqrt{\frac{2(K_{sr_i} + 1)\gamma}{\xi_{sr_i}}}\right) f_{\xi_{rd}}(\gamma) + Q_1\left(\sqrt{2K_{rd}}, \sqrt{\frac{2(K_{rd} + 1)\gamma}{\xi_{rd}}}\right) f_{\xi_{sr_i}}(\gamma) \quad (11)$$

The LT of the random variable $\gamma_{ub} = \gamma_{sd} + \xi_{sum}$ is obtained by using IVT at high SNR regime. Since $\gamma \rightarrow 0$ as $\gamma_0 \rightarrow \infty$, we can write

$$\lim_{s \rightarrow \infty} s\mathcal{L}(f_{\gamma_{sd}}(\gamma)) = \lim_{\gamma \rightarrow \infty} f_{\gamma_{sd}}(\gamma) \quad (12)$$

This implies

$$\mathcal{L}(f_{\gamma_{sd}}(\gamma)) = \frac{1}{s} f_{\gamma_{sd}}(0) \quad (13)$$

Similarly the LT of the PDF of random variable ξ_{sum} can be expressed as

$$\mathcal{L}(f_{\xi_{sum}}(\gamma)) = \frac{1}{s^M} \prod_{i=1}^M f_{\xi_{min,i}}(0) \quad (14)$$

Now by using the multiplication properties of LT, the LT of the PDF of random variable γ_{ub} for i.n.d fading channel can be written as

$$\mathcal{L}(f_{\gamma_{ub}}(\gamma)) = \frac{1}{s^{M+1}} f_{\gamma_{sd}}(0) \prod_{i=1}^M f_{\xi_{min,i}}(0) \quad (15)$$

The PDF of random variable γ_{ub} is obtained by applying inverse LT (ILT) on the above

$$f_{\gamma_{ub}}(\gamma) = \frac{1}{M!} \gamma^M f_{\gamma_{sd}}(0) \prod_{i=1}^M f_{\xi_{min,i}}(0) \quad (16)$$

By integrating the above and substituting the value of $f_{\gamma_{sd}}(0)$ and $f_{\xi_{min,i}}(0)$, the outage probability can be expressed as

$$p_{out} = \frac{1}{(M+1)! \bar{\gamma}_{sd}} \prod_{i=1}^M \left(\frac{(K_{sr_i} + 1)}{\xi_{sr_i}} e^{K_{sr_i}} + \frac{(K_{rd} + 1)}{\xi_{rd}} e^{K_{rd}} \right) \gamma^{M+1} \quad (17)$$

3.2 Asymmetric channel II:

For asymmetric channel II, signal in S-D and R-D links experience Rayleigh distribution and signal in S-R link experiences Rician distribution. For this scenario, we use $\xi_{sd} = P_s |h_{sd}|^2$, $\xi_{sr_i} = P_s |h_{sr_i}|^2$ and $\gamma_{rd} = P_s |h_{rd}|^2$. The end-to-end instantaneous SNR can be expressed as

$$\gamma_{ub} = \gamma_0 \xi_{sd} + \gamma_0 \xi_{sum} \quad (18)$$

where $g_{sum} = \sum_{i=1}^M g_{min,i}$ and $g_{min,i} = \min(\xi_{sr_i}, \gamma_{r_i,d})$. Similarly as the previous section, the PDF of $g_{min,i}$ is expressed as

$$f_{g_{min,i}}(\gamma) = Q_1 \left(\sqrt{2K_{sr_i}} \sqrt{\frac{2(K_{sr_i} + 1)\gamma}{\xi_{sr_i}}} \right) f_{\gamma_{r_i,d}}(\gamma) + f_{\xi_{sr_i}}(\gamma) (1 - F_{\gamma_{r_i,d}}(\gamma)) \quad (19)$$

where $F_{\gamma_{r_i,d}}(\gamma)$ is the CDF of the random variable $\gamma_{r_i,d}$.

Similarly as previous, the PDF of γ_{ub} for this asymmetric channel can be derived as

$$f_{\gamma_{ub}}(\gamma) = \frac{1}{M!} \gamma^M f_{\xi_{sd}}(0) \prod_{i=1}^M f_{g_{min,i}}(0) \quad (20)$$

By integrating (20), the outage probability for the asymmetric channel II can be expressed as

$$P_{out} = \frac{1}{(M+1)!} \frac{(K_{sd} + 1)}{\xi_{sd}} e^{K_{sd}} \prod_{i=1}^M \left(\frac{(K_{sr_i} + 1)}{\xi_{sr_i}} e^{K_{sr_i}} + \frac{1}{\gamma_{r_i,d}} \right) \gamma^{M+1} \quad (21)$$

4. Opportunistic AF relaying

In this section, we analyze the outage probability of opportunistic AF relaying over the same asymmetric fading scenario. For the relay selection, we use maximum SNR approach provided in Bletsas et al. (2007). The equivalent instantaneous end-to-end SNR for opportunistic AF relaying is given as Zhao et al. (2006)

$$\gamma = \frac{P_s |h_{sd}|^2}{N_{sd}} + \max_{i=\{1,2,\dots,M\}} \frac{\frac{P_s |h_{sr_i}|^2}{N_{sr_i}} \frac{P_s |h_{r_i,d}|^2}{N_{r_i,d}}}{\frac{P_s |h_{sr_i}|^2}{N_{sr_i}} + \frac{P_s |h_{r_i,d}|^2}{N_{r_i,d}} + 1} \quad (22)$$

The upper bound of instantaneous SNR for the above can be written as

$$\gamma_{max} = P_s |h_{sd}|^2 \gamma_0 + \max_{i=\{1,2,\dots,M\}} \min(P_s |h_{sr_i}|^2 \gamma_0, P_s |h_{r_i,d}|^2 \gamma_0) \quad (23)$$

4.1 Asymmetric channel I:

In asymmetric channel I, the outage performance can be expressed as

$$p_{out} = Pr[\gamma_{ub} < \gamma] \quad (24)$$

where $\gamma_{ub} = \gamma_{max}/\gamma_0$, $\gamma = (2^{2R} - 1)/\gamma_0$, $\xi_{max} = \max(\xi_{min,1}, \xi_{min,2}, \dots, \xi_{min,M})$ and $\xi_{min,i} = \min(\xi_{sr_i}, \xi_{r_i,d})$. The CDF of the random variable ξ_{max} for i.n.d fading channel can be expressed as

$$F_{\xi_{max}}(\gamma) = \prod_{i=1}^M F_{\xi_{min,i}}(\gamma) \quad (25)$$

and the corresponding PDF of ξ_{max} is obtained as

$$f_{\xi_{max}}(\gamma) = \sum_{i=1}^M f_{\xi_{min,i}}(\gamma) \prod_{\substack{j=1 \\ j \neq i}}^M f_{\xi_{min,j}}(\gamma) \quad (26)$$

Since $F_{\xi_{min,i}}(0) = 0$, the $(M-1)^{th}$ order derivative of (26) at high SNR, i.e., at $\gamma = 0$ for $\gamma_0 \rightarrow \infty$, can be derived as

$$\frac{\partial^{M-1}}{\partial \gamma^{M-1}} f_{\xi_{max}}(\gamma) \Big|_{\gamma=0} = M! \prod_{i=1}^M f_{\xi_{min,i}}(0) \quad (27)$$

By using LT of M^{th} order differentiation, we can write

$$\mathcal{L}\left(\frac{\partial^{M-1}}{\partial \gamma^{M-1}} f_{\xi_{max}}(\gamma)\right) = s^{M-1} \mathcal{L}(f_{\gamma_{max}}(\gamma)) - s^{M-2} f_{\gamma_{max}}(0) - \dots - f_{\gamma_{max}}^{(M-2)}(0) \quad (28)$$

Since $f_{\gamma_{max}}(0) = f_{\gamma_{max}}^{(1)}(0) = \dots = f_{\gamma_{max}}^{(M-2)}(0) = 0$, by using the IVT of LT, we can write

$$\mathcal{L}(f_{\gamma_{max}}(\gamma)) \Big|_{\lim s \rightarrow \infty} = \frac{1}{s^M} \frac{\partial^{M-1}}{\partial \gamma^{M-1}} f_{\xi_{max}}(\gamma) \Big|_{\gamma=0} \quad (29)$$

The LT of the PDF of random variable $\gamma_{ub} = \gamma_{sd} + \xi_{max}$ over i.n.d can be written as

$$\begin{aligned} \mathcal{L}(f_{\gamma_{ub}}(\gamma)) &= \mathcal{L}(f_{\gamma_{sum}}(\gamma)) \mathcal{L}(f_{\gamma_{max}}(\gamma)) \\ &= \frac{1}{s^{M+1}} f_{\gamma_{sd}}(0) \frac{\partial^{M-1}}{\partial \gamma^{M-1}} f_{\xi_{max}}(\gamma) \Big|_{\gamma=0} \\ &= \frac{M!}{s^{M+1}} f_{\gamma_{sd}}(0) \prod_{i=1}^M f_{\xi_{min,i}}(0) \end{aligned} \quad (30)$$

with respect to s , the PDF of γ_{ub} is obtained by applying the ILT on the above as

$$f_{\gamma_{ub}}(\gamma) = \gamma^M f_{\gamma_{sd}}(0) \prod_{i=1}^M f_{\xi_{min,i}}(0) \quad (31)$$

The corresponding outage probability or CDF of γ_{ub} is obtained by integrating the above as.

$$p_{out} = \frac{1}{(M+1)\bar{\gamma}_{sd}} \prod_{i=1}^M \left(\frac{K_{sr_i} + 1}{\bar{\xi}_{sr_i} e^{K_{sr_i}}} + \frac{K_{r_i d} + 1}{\bar{\xi}_{r_i d} e^{K_{r_i d}}} \right) \gamma^{M+1} \quad (32)$$

4.2 Asymmetric channel II:

Similarly, in asymmetric channel II, the outage performance can be expressed as

$$p_{out} = Pr[\gamma_{ub} < \gamma] \quad (33)$$

where $\gamma_{ub} = \xi_{sd} + g_{max}$, $g_{max} = \max(g_{min,1}, g_{min,2}, \dots, g_{min,M})$ and $g_{min,i} = \min(\gamma_0 \xi_{sr_i}, \gamma_0 \gamma_{r,d})$.

As the previously, the LT of the random variable of γ_{ub} over i.n.d fading channel can be written as

$$\mathcal{L}(f_{\gamma_{ub}}(\gamma)) = \frac{M!}{s^{M+1}} f_{\gamma_{sd}}(0) \prod_{i=1}^M f_{g_{min,i}}(0) \quad (34)$$

The PDF of γ_{ub} is obtained by applying the ILT on the above as

$$f_{\gamma_{ub}}(\gamma) = \gamma^M f_{\gamma_{sd}}(0) \prod_{i=1}^M f_{g_{min,i}}(0) \quad (35)$$

The corresponding outage probability or CDF of γ_{ub} is obtained by integrating the above as

$$p_{out} = \frac{1}{(M+1)\bar{\gamma}_{sd}} \prod_{i=1}^M \left(\frac{K_{sr_i} + 1}{\xi_{sr_i} e^{K_{sr_i}}} + \frac{1}{\bar{\gamma}_{r,d}} \right) \gamma^{M+1} \quad (36)$$

5. Numerical

In this section, analytical and Monte-Carlo simulation results are presented. Since the channel are i.n.d, we set different means for different S-R_i/R_i-D links. In the Rician fading channel, the Rician factor K_{ab} is uniformly distributed in [2,3] and the mean $\bar{\gamma}_{ab}$ of NLOS components are uniformly distributed in [0,1]. The LOS components are derived for a given value of K_{ab} and $\bar{\gamma}_{ab}$. The number of relay nodes is set to 2, 4, 5 and 6.

Fig. 2, Fig. 3 and Fig. 4 show the lower bound of outage probability of repetition-based AF relaying over asymmetric channel I, asymmetric channel II and a comparison among the different fading channels. Since the analytical outage probability are derived based on high SNR assumption, analytical results converge with Monte-Carlo simulation results at high SNR value. From Fig. 2 and Fig. 3, it is clear that the diversity of repetition-based AF relaying increases with the number of relay nodes.

From Fig. 4, it is clear that the outage performance over Rician fading channel outperforms all other fading scenarios due to the presence of LOS signal. On the other hand, due to the absence of LOS signals, Rayleigh fading channel has poorer outage performance than all other fading scenarios. When the outage performance is compared between two different asymmetric channels, asymmetric channel I provides better outage performance than the asymmetric channel II. It is because of S-R link experience LOS signal in asymmetric channel I, so, there is less chance to amplify the noise by relay nodes. However, for the asymmetric channel II, S-R is a NLOS situation, there is more chance to amplify the noise by relay nodes and send it to the destination.

Fig. 5, Fig. 6 and Fig. 7 show the lower bound of outage probability of opportunistic AF relaying over asymmetric channel I, asymmetric channel II and a comparison among the different fading channels. As similar as repetition-based relaying, analytical outage performance converges with Monte-Carlo simulation results at high SNR values. In opportunistic relaying, the performance as well as the diversity increase with the number of

relay nodes. When the outage performance is compared among the fading channel, opportunistic relaying shows the same characteristic as repetition based relaying. Without providing any extra simulation, it is easily concluded that opportunistic AF relaying provides better outage performance than the repetition-based AF relaying.

6. Conclusions

This work investigates the outage performance of repetition-based and opportunistic AF relaying over two different asymmetric fading channel. The lower bound of outage probability is derived for high SNR regime and validated through the Monte-Carlo simulation studies. It is observed that asymmetric channel I has better outage performance than that of asymmetric channel II for both the repetition-based and opportunistic AF relaying, and opportunistic AF relaying provides better outage performance than the repetition-based AF relaying.

7. Acknowledgments

The authors would like to thank the European IST-FP7 WHERE project for support of this work.

8. References

- Bletsas, A., Shin, H. and Win, M. Z. (2007). Cooperative communication with outage optimal opportunistic relaying, *IEEE Transactions on Wireless Communications* 6: 3450–3459.
- Hwang, K.-S., Ko, Y.-C. and Alouini, M.-S. (2007). Outage probability of cooperative diversity systems with opportunistic relaying based on decode-and-forwards, *IEEE Transactions on Wireless Communications* 7: 5100–5106.
- Katz, M. and Shamai, S. (2009). Relaying protocols for two colocated users, *IEEE Transactions on Information Theory* 52: 2329 – 2344.
- Krikidis, I. and Thompson, J. (2008). Amplify-and-Forward with partial relay selection, *IEEE Communications Letters* 12: 235–237.
- Laneman, J. N., Tse, D. N. C. and Wornell, G. W. (2004). Cooperative diversity in wireless networks: Efficient protocols and outage behavior, *IEEE Transactions of Information Theory* 50: 3062–3080.
- Liu, K. J. R., Sadek, A. K., Su, W. and Kwasinski, A. (2009). *Cooperative Communications and Networking*, Cambridge.
- Michalopoulos, D. and Karagiannidis, G. (2008). Performance analysis of single relay selection in Rayleigh fading, *IEEE Transactions on Wireless Communications* 7(10): 3718–3724.
- Nosratinia, A., Hunter, T. E. and Hedayat, A. (2004). Cooperative communication in wireless networks, *IEEE Communications Magazine* 42: 74–80.
- Paulraj, A., Gore, D., Nabar, R. and Bolcskei, H. (2004). An overview of mimo communications - a key to gigabit wireless, *Proceedings of the IEEE* 92(2): 198–218.
- Savazzi, S. and Spagnolini, U. (2008). Cooperative fading regions for decode and forward relaying, *IEEE Transactions on Information Theory* 54(11): 4908–4924.

- Suraweera, H., Karagiannidis, G. and Smith, P. (2009). Performance analysis of the dualhop asymmetric fading channel, *IEEE Transactions on Wireless Communications Letters* 8: 2783–2788.
- Suraweera, H., Louie, R., Li, Y., Karagiannidis, G. and Vucetic, B. (2009). Two hop amplify-and-forward transmission in mixed Rayleigh and Rician fading channels, *IEEE Communications Letters* 13(4): 227–229.
- Vicario, J., Bel, A., Lopez-Salcedo, J. and Seco, G. (2009). Opportunistic relay selection with outdated csi: outage probability and diversity analysis, *IEEE Transactions on Wireless Communications* 8(6): 2872–2876.
- Xu, F., Lau, F. C. M., Zhou, Q. F. and You, D. W. (2009). Outage performance of cooperative communication systems using opportunistic relaying and selection combining receiver, *IEEE Signal Processing Letters* 16: 113–116.
- Zhao, Y., Adve, R. and Lim, T. (2007). Improving amplify-and-forward relay networks: optimal power allocation versus selection, *IEEE Transactions on Wireless Communications* 6(8): 3114–3123.
- Zhao, Y., Adve, R. and Lim, T. J. (2005). Outage probability at arbitrary SNR with cooperative diversity, *IEEE Communications Letters* 9: 700–703.
- Zhao, Y., Adve, R. and Lim, T. J. (2006). Symbol error rate of selection Amplify-and-Forward relay systems, *IEEE Communications Letters* 10: 757–759.
- Zhu, Y., Xin, Y. and Kam, P.-Y. (2008). Outage probability of Rician fading relay channels, *IEEE Transactions on Vehicular Technology* 57(4): 2648–2652.
- Zou, Y., Zheng, B. and Zhu, J. (2009). Outage analysis of opportunistic cooperation over Rayleigh fading channels, *IEEE Transactions on Wireless Communications* 8(6): 3077–3085.

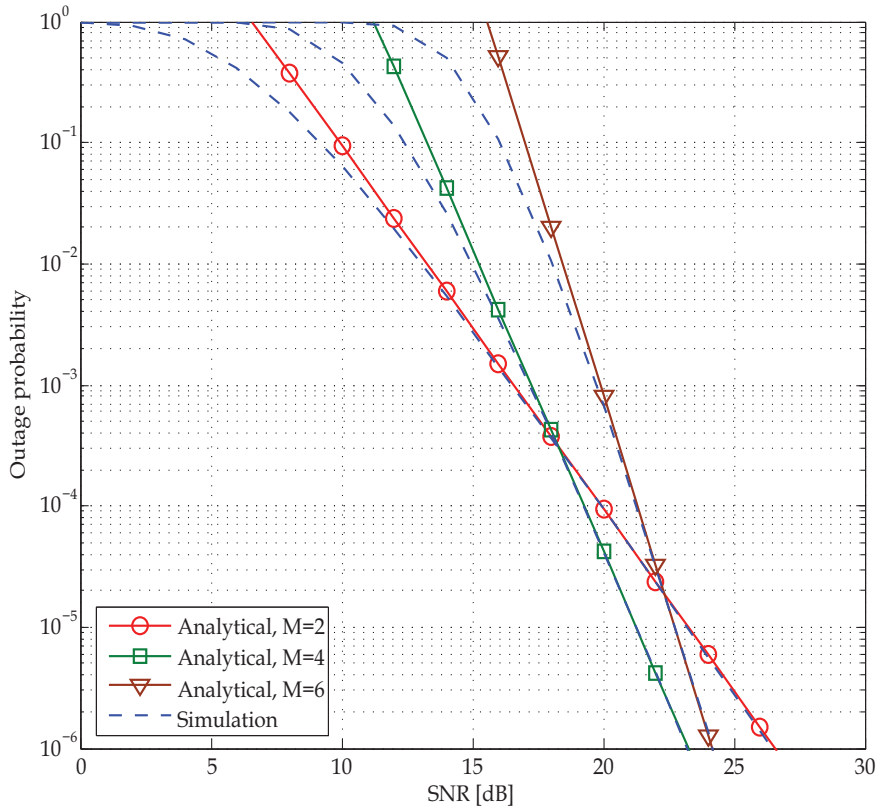


Fig. 2. The outage probability of repetition-based AF relaying over asymmetric channel I. The number of relay node is selected $M = 2$, $M = 4$ and $M = 6$.

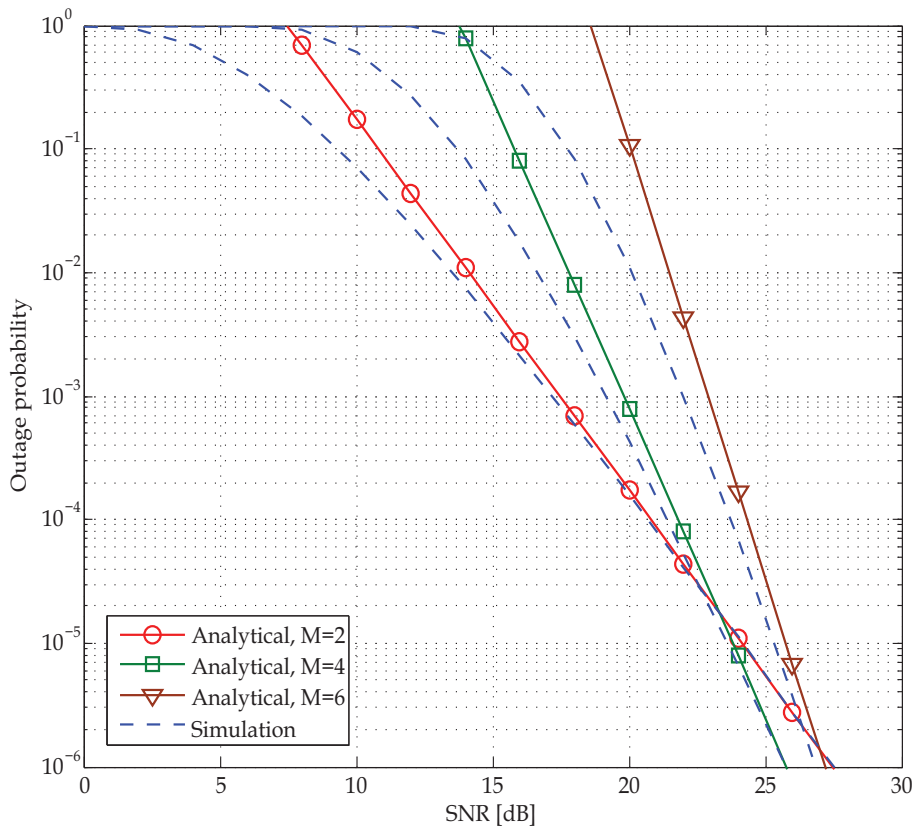


Fig. 3. The outage probability of repetition-based AF relaying over asymmetric channel II. The number of relay node is selected $M = 2$, $M = 4$ and $M = 6$.

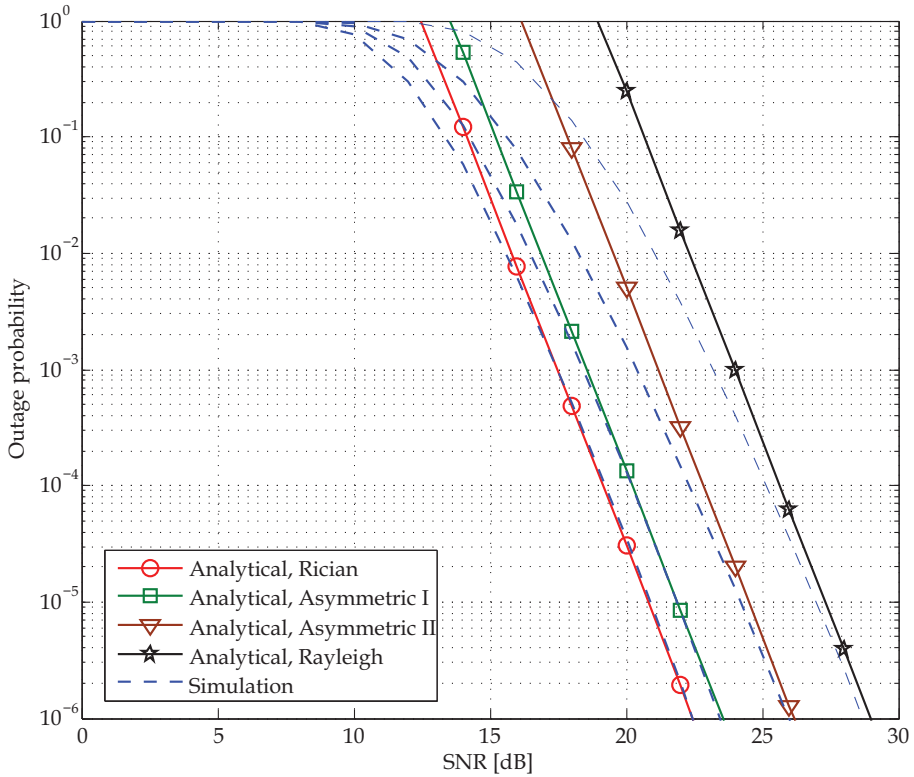


Fig. 4. The comparison of outage probability of repetition-based relaying over different fading channel such as Rician fading, Rayleigh fading, asymmetric channel I and asymmetric channel II. The number of relay nodes is of $M = 5$.

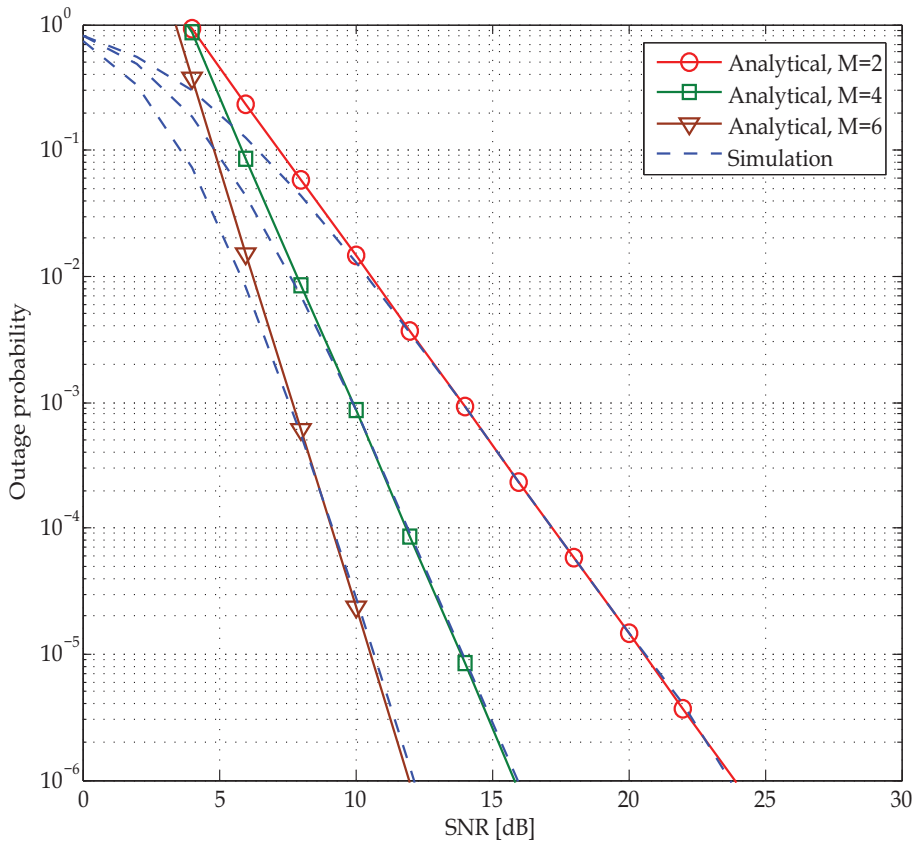


Fig. 5. The outage probability of opportunistic AF relaying over asymmetric channel I. The number of relay node is selected $M = 2$, $M = 4$ and $M = 6$.

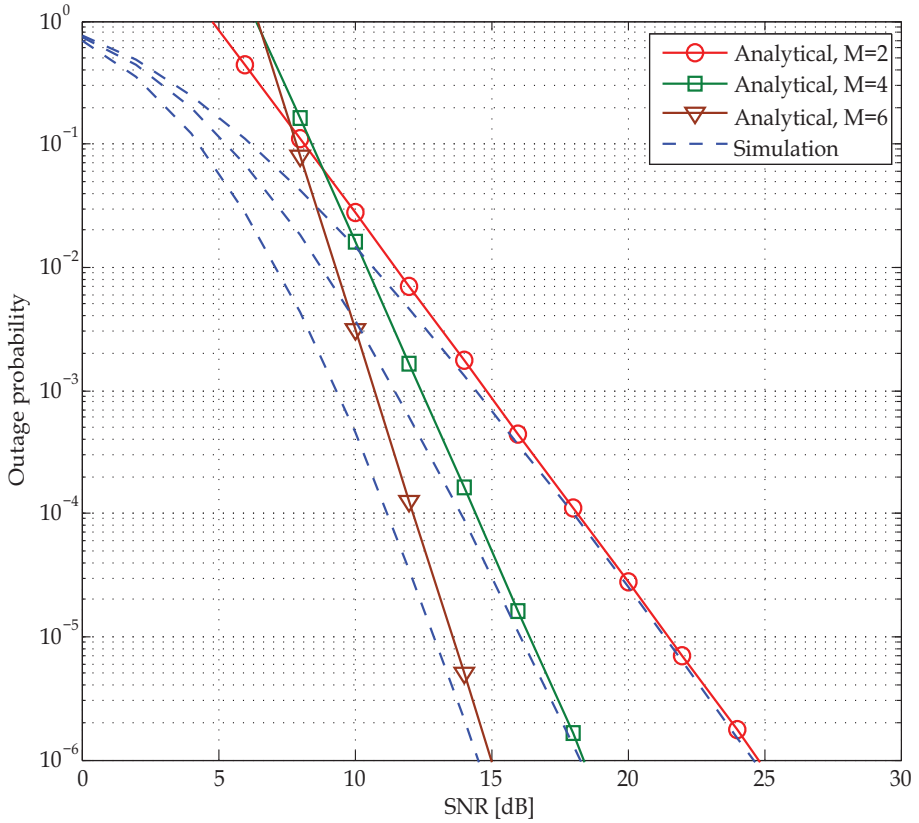


Fig. 6. The outage probability of opportunistic AF relaying over asymmetric channel II. The number of relay node is selected $M = 2$, $M = 4$ and $M = 6$.

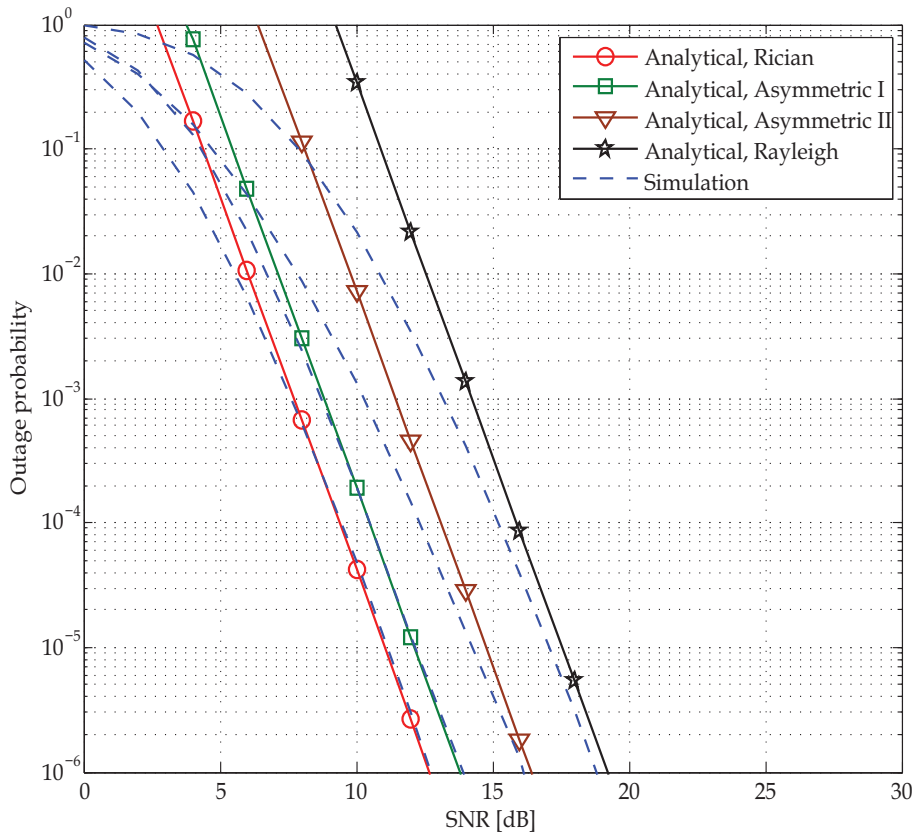


Fig. 7. The comparison of outage probability of opportunistic relaying over different fading channel such as Rician fading, Rayleigh fading, asymmetric channel I and asymmetric channel II. The number of relay nodes is of $M = 5$.

Indoor Radio Network Optimization

Lajos Nagy

*Department of Broadband Communications and Electromagnetic Theory
Budapest University of Technology and Economics
Hungary*

1. Introduction

The new focus of wireless communication is shifting from voice to multimedia services. User requirements are moving from underlying technology to the simply need reliable and cost effective communication systems that can support anytime, anywhere, any device. The most important trends in global mobile data traffic forecast are:

Globally, mobile data traffic will double every year through 2014, increasing 39 times between 2009 and 2014.,

Almost 66 percent of the world's mobile data traffic will be video by 2014. (Cisco, 2010)

While a significant amount of traffic will migrate from mobile to fixed networks, a much greater amount of traffic will migrate from fixed to mobile networks. In many countries mobile operators are offering mobile broadband services at prices and speeds comparable to fixed broadband. Though there are often data caps on mobile broadband services that are lower than those of fixed broadband, some consumers are opting to forgo their fixed lines in favor of mobile.

There is a growing interest in providing and improving radio coverage for mobile phones, short range radios and WLANs inside buildings. The need of such coverage appears mainly in office buildings, shopping malls, train stations where the subscriber density is very high. The cost of cellular systems and also the one of indoor wireless systems depend highly on the number of base stations required to achieve the desired coverage for a given level of field strength. (Murch 1996)

The other promising technique is the Hybrid Fiber Radio (HFR)-WLAN which is combines the distribution and radio network. The advantages of using analogue optical networks for delivering radio signals from a central location to many remote antenna sites have long been researched and by using the high bandwidth, low loss characteristics of optical fiber, all high frequency and signal processing can be performed centrally and transported over the optical network directly at the carrier frequency. The remote site simplicity makes possible the network cheap and simple, requiring only optoelectronic conversion (laser diodes and photo-detectors), filtering and amplification. Such Remote Units (RU) would also be cheap, small, lightweight, and easy to install with low power consumption.

The design objectives can list in the priority order as RF performance, cost, specific customer requests, ease of installation and ease of maintenance. The first two of them are close related to the optimization procedure introduced and can take into account at the design phase of the radio network.

There are already numerous optimization methods published which can be applied to the optimal design of such indoor networks (Wu 2007, Adickes 2002, Portilla-Figueras 2009, Pujji 2009). The recently published methods use any heuristic technique for finding the optimal Access Point (AP) or RU positions. Common drawback of the methods are the slow convergence in a complex environment like the indoor one because all of the methods are using the global search space i.e. the places for AP-s are searched globally.

This chapter presents approaches in optimizing the indoor radio coverage using multiple access points for indoor environments. First the conventional Simple Genetic Algorithm (SGA) is introduced and used to determine the optimal access point positions to achieve optimum coverage. Next to overcome the disadvantage of SGA two optimization methods are applied Divided Rectangles (DIRECT) global optimization technique and a new hierarchic optimization method is introduced and comparisons are made for the methods deployed.

The main advantage of the proposed method is the reduction of the search space by using two step procedure starting with simple radio propagation method based AP position estimation and thereafter heuristic search using Motley Keenan radio propagation method with heuristic search.

2. Hybrid fiber radio architecture

Microwave radio-frequency transport over fibre, is an already widely used approach which allows the radio functionality of several Base Stations (BS) to be integrated in a centralised headend unit (Schuh, 1999).

Moreover, it offers fixed and mobile wireless broadband access with a radio-independent fibre access network. Different radio feeder concepts such as Intermediate Frequency (IF) over fibre with electrical frequency conversion at the RAU or direct Radio Frequency (RF) transport are possible.

Few existing Hybrid Fiber Radio interfaces are

DECT - narrowband access for indoor multi-cell cordless telephony, with indoor range from 20 up to 50 metres, and for outdoor Wireless Local Loop (WLL) with a radio range up to a few kilometres.

GSM cellular mobile system provides narrowband access for speech and data services. Typical indoor DCS-1800 cell radius is from about 10 to 50 m and outdoor cell radius for GSM-900/DCS-1800 vary often between 50 to 1000 m.

W-LANs (IEEE 802.11) operate in 80 MHz of spectrum using the 2.4 GHz ISM band, giving indoor access originally designed to high data rates, up to 2 Mbit/s, with coverage areas up to 250 m.

UMTS will operate at ~2 GHz with up to 60 MHz of spectrum. It can provide features like 2nd generation mobile systems but will also offer multimedia services like video telephony, up to 2 Mbit/s for low mobility. Supported cell sizes for indoor applications are up to ~100 metres, and for outdoor applications cell size can be up to a few tens of kilometres (suburban areas), by supporting different mobility features. UMTS will be a public operated system.

One possible application of the HFR network is using analog optical links to transmit modulated RF signals. It serves to transmit the RF signals down- and uplink, i.e. to and from central units (CU) to base stations (BS) called also radio ports. Basic design is shown in Fig. 1, using wavelength duplex fiber star (T1) and fiber bus (T2) topology. This technique is the mostly used one in cellular HFR networks. [1,6]

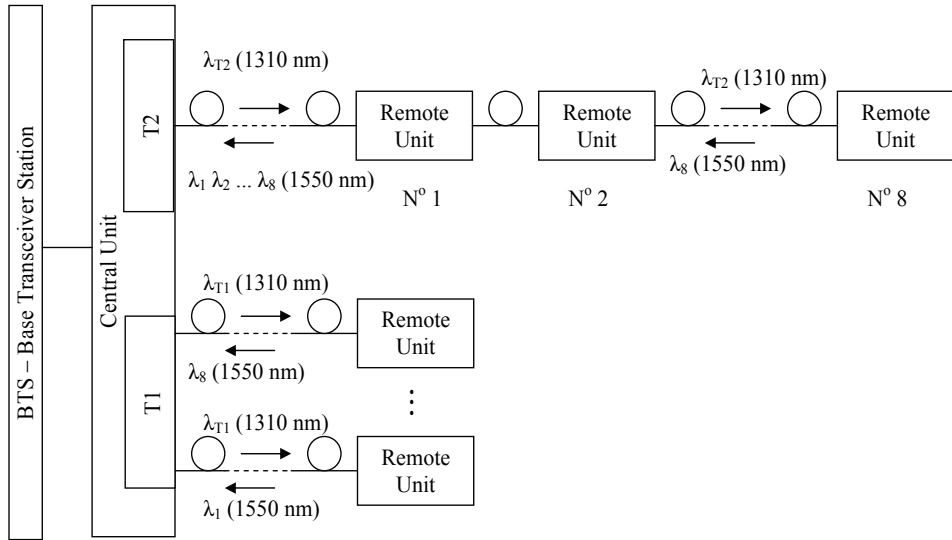


Fig. 1. HFR cellular architecture using one fiber star (T1) and one fiber bus (T2) topology

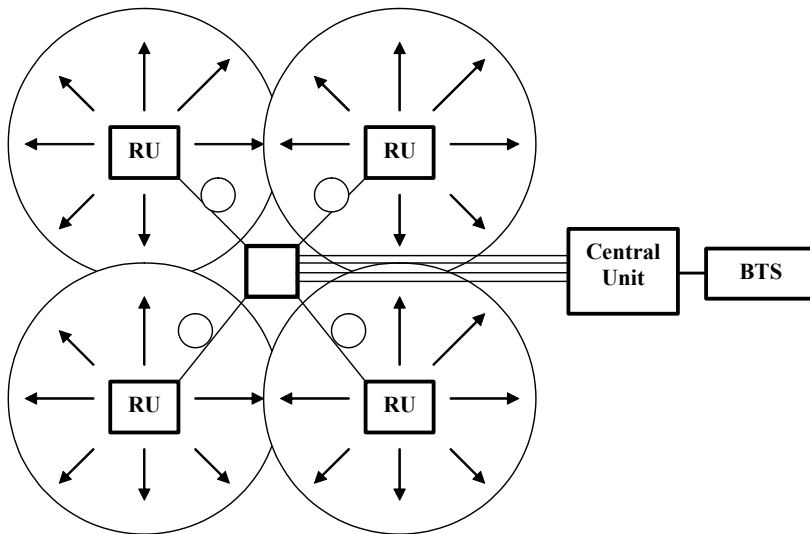


Fig. 2. HFR cellular architecture with Remote Units and Central Units

The other technique uses direct modulation of laser diode and more suitable for WLAN applications. The Fig. 3. shows the combination of IEEE 802.11a and 11.g WLAN services using HFR technology.

The main parts of the HFR network in Fig. 3. are the Local Transceiver Unit with circulator, electro-optical converters and the Remote Unit with electro-optical converters, antennas. The IEEE 802.11a and 11.g WLAN access points are used in unchanged form accessing to the wired internet network.

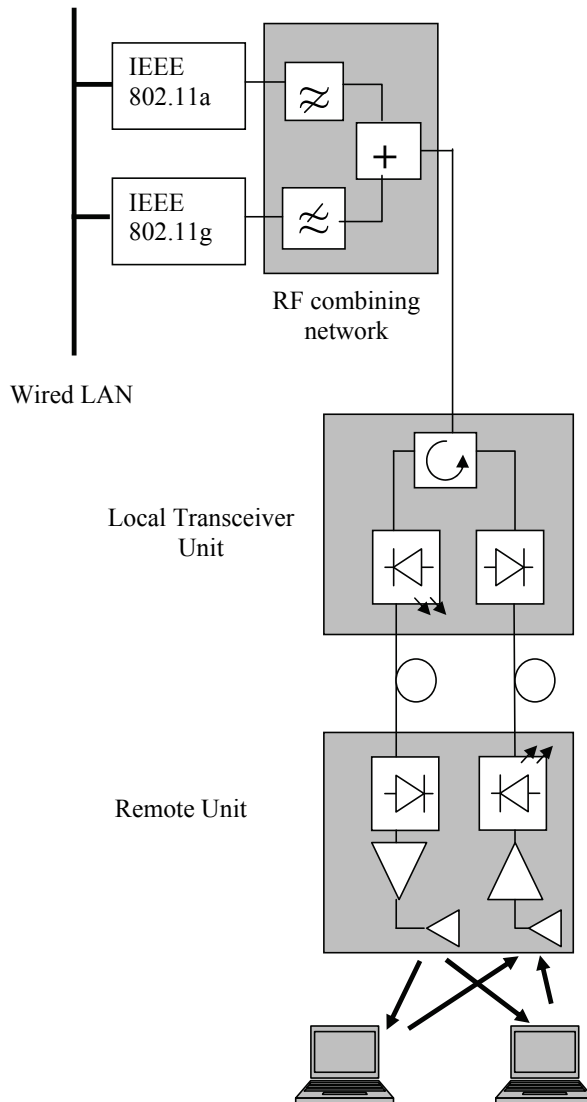


Fig. 3. HFR WLAN architecture

The indoor radio coverage of HFR network is basically determined by the RU positions. There are many factors on choosing these positions such as RF performance, cost, specific customer requests, ease of installation, ease of maintenance, but the optimal radio coverage achievable by a minimum number of the RUs is the most important one.

The next parts introduces the radio propagation modeling used in indoor environment and the optimization method for determining optimum RU positions for best radio coverage which is usually the main aim of the wireless design but the optimization method proposed can be easily amended of further objectives.

3. The indoor radiowave propagation model and the building database

In our article the Motley-Keenan (Keenan & Motley, 1990) model was used to analyze indoor wave propagation. This empirical type prediction model based on considering the influence of walls, ceilings and floors on the propagation through disparate terms in the expression of the path loss.

The overall path loss according to this model can be written as

$$L = L_F + L_a \quad (1)$$

where L_F is the free space path loss and L_a is an additional loss expressed as

$$L_a = L_c + \sum_{i=1}^I k_{wi} L_{wi} + \sum_{j=1}^J k_{fj} L_{fj} \quad (2)$$

where L_c is an empirical constant term, k_{wi} is the number of penetrated i type walls, k_{fj} is the number of penetrated floors and ceilings of type j , I is the number of wall types and J is the number of floor and ceiling types.

For the analyzed receiver position, the numbers k_i and k_j have to be determined through the number of floors and walls along the path between the transmitter and the receiver antennas. In the original paper (Keenan & Motley, 1990) only one type of walls and floors were considered, in order for the model to be more precise a classification of the walls and floors is important. A concrete wall for example could present very varying penetration losses depending on whether it has or not metallic reinforcement.

It is also important to state that the loss expressed in (Eq. 2) is not a physical one, but rather model coefficients, that were optimized from measurement data. Constant L_c is the result of the linear regression algorithm applied on measured wall and floor losses. This constant is a good indicator of the loss, because it includes other effects also, for example the effect of furniture.

For the considered office type building, the values for the regression parameters have been found. (Table 1.)

The Motley-Keenan model regression parameters have been determined using Ray Launching deterministic radiowave propagation model. These calculations were made for the office-type building floor of the Department of Broadband Infocommunication and Electromagnetic Theory at Budapest University of Technology and Economics (Fig. 4.). The frequency was chosen to 2450 MHz with a $\lambda/2$ transmitter dipole antenna mounted on the 3m height ceiling at the center of the floor.

The receiver antenna has been applied to evaluate the signal strength at $(80 \times 5) \times (22 \times 5) = 44000$ different locations in the plane of the receiver. At each location the received signal strength was obtained by RL method using ray emission in a resolution of 1° . A ray is followed until a number of 8 reflections are reached and the receiver resolution in pixels has an area of $0.2 \times 0.2 \text{ m}^2$. The receiver plane was chosen at the height of 1.2 m.

The wall construction is shown on Fig. 4. made of primarily brick and concrete with concrete ceiling and floor, the doors are made of wood. The coefficients of the model have been optimized on the data gathered by the RL simulation session described above.

The floor view and polygonal partitioning is shown on Fig. 5., which is based on the concept described next.

Wall type	Nr. of Layers	Layer widths	Regression parameter [dB]
Brick	1	Brick - 6 cm	4.0
Brick	1	Brick - 10 cm	5.58
Brick	1	Brick - 12 cm	6.69
Brick+ Concrete	3	Brick - 6 cm Concrete - 20 cm Brick - 6 cm	11.8
Brick+ Concrete	3	Brick - 10 cm Concrete - 12 cm Brick - 10 cm	14.8
Brick+ Concrete	3	Brick - 6 cm Concrete - 10 cm Brick - 6 cm	9.3
Brick	1	Brick - 15 cm	8.47
Concrete	1	Concrete - 15 cm	6.56
Concrete	1	Concrete - 20 cm	8
Concrete	3	Concrete - 15 cm Air - 2 cm Concrete - 15 cm	12.47
Glass	3	Glass - 3 mm Air - 10 cm Glass - 3 mm	0
Plasterboard	1	Plasterboard - 5 cm	4.5
Wood	1	Wood - 6 cm	0.92
Wood	1	Wood - 10 cm	0.17

Table 1. The regression parameters

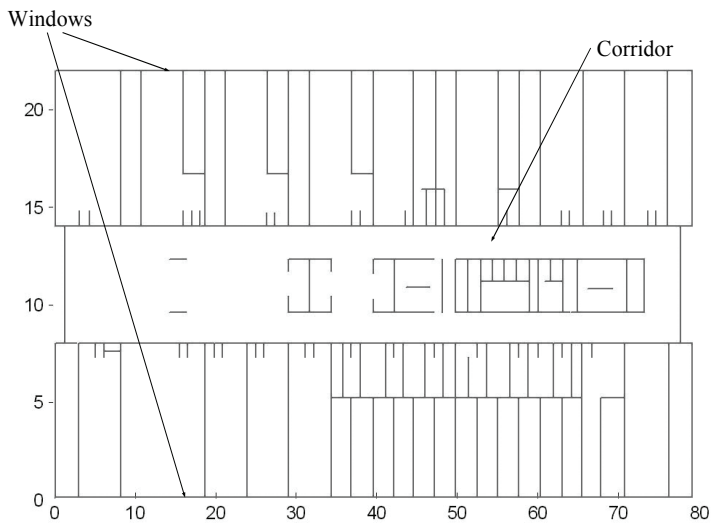


Fig. 4. The building database



Fig. 5. Floor view and polygon data base of V2 building at BUTE

The geometrical description of the indoor scenario is based on the same concept that the walls has to be partitioned to surrounding closed polygons and every such polygons are characterized by its electric material parameters.

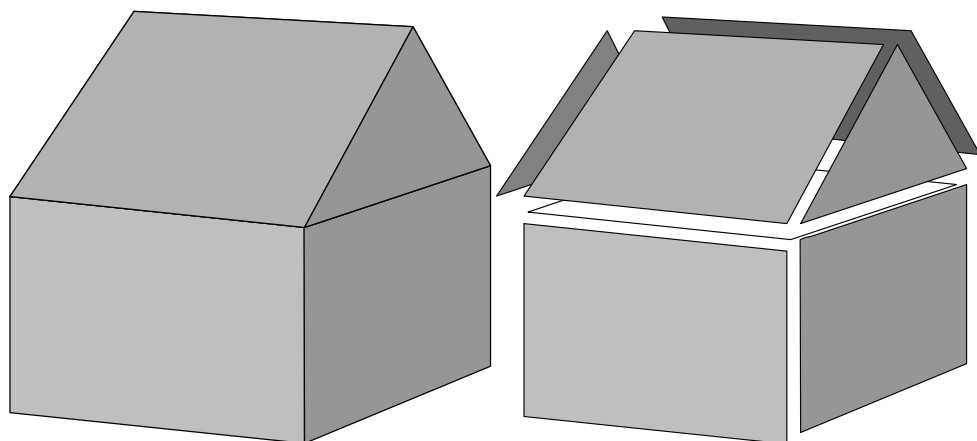


Fig. 6. Polygon representation of building structure

The data base for the ray tracing method in our applications can not contain cut-out surfaces directly, such as windows, doors. Therefore the cut-out surface description is based on surface partitioning of the geometry as can be seen in Fig. 7.

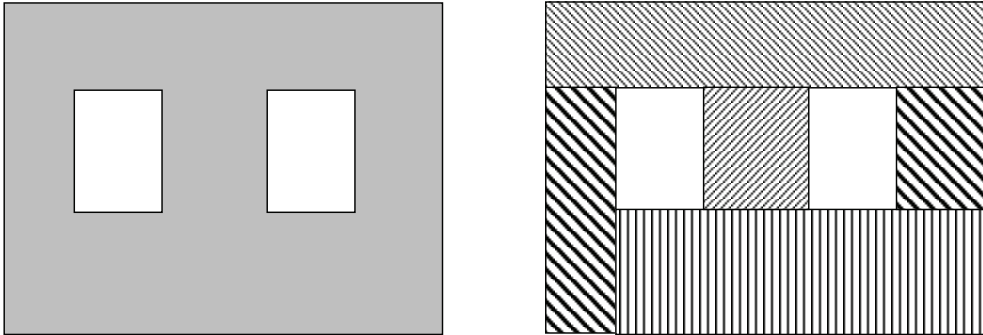


Fig. 7. A possible polygonal partitioning of windowed walls for ray tracing method

4. Optimization methods

There are already numerous optimization methods published which can be applied to the optimal design of such Hybrid Fiber Radio indoor networks (Wu 2007, Adickes 2002, Portilla-Figueras 2009, Pujji 2009). The recently published methods use any heuristic technique for finding the optimal Access Point (AP) or RU positions. Common drawback of the methods are the slow convergence in a complex environment like the indoor one because all of the methods are using the global search space i.e. the places for AP-s are searched globally.

Heuristic search and optimization is an approach for solving complex and large problems that overcomes many shortcomings of traditional (gradient type) optimization techniques. Heuristic optimization techniques are general purpose methods that are very flexible and can be applied to many types of objective functions and constraints. Another advantage of heuristic methods is their simplicity because of its gradient-free nature. Gradient free optimization methods are primarily based on the objective function values and are suitable for problems either with many parameters or with computationally expensive objective functions.

In the paper two global optimization methods the Simple Genetic Algorithm (SGA) and a method using Divided Rectangles (DIRECT) global search algorithm are used with wave propagation solver as can be seen in Fig. 8.

4.1 Optimization method through Simple Genetic Algorithms (SGA)

Genetic Algorithms (GA) are increasingly being applied to complex problems. Genetic Algorithm optimizers are robust, stochastic search methods modeled on the principles and concepts of natural selection. (Nagy 2000, Farkas 2001, Michielssen 1999, Michalewicz 1996) Genetic Algorithms (GA) are increasingly being applied to difficult optimization problems. GA optimizers are robust, stochastic search methods modeled on the principles and concepts of natural selection.

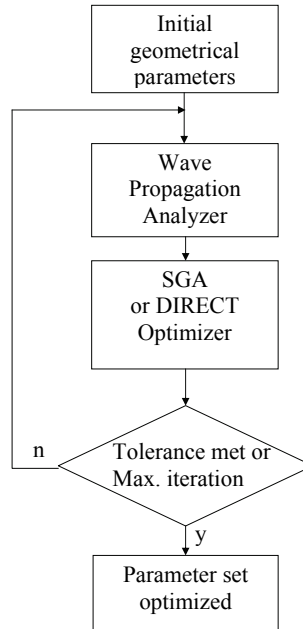


Fig. 8. Diagram of Wave Propagation analyzer and optimizer

If a receiver position that is fully described by N_{par} parameters arranged in a vector $x = \{x_i \mid i=1, \dots, N_{par}\}$ is considered, then the knowledge of x permits the evaluation of the objective function $f(x)$, which indicates the worth of a design (the area coverage percentage). It is assumed that x_i take on either real or discrete values, and that $f(x)$ needs to be maximized.

The GA does not operate on x but on a discrete representation or chromosome $p = \{g_i \mid i=1, \dots, N\}$ of x , each parameter x_i being described by a gene g_i . Each gene g_i in turn consists of a set of N_{all}^i all that are selected from a finite alphabet and that together decode a unique x_i .

The GA does not limit themselves to the iterative refinement of a single coded design candidate; instead the simple GA (SGA) simultaneously acts upon a set of candidates or population

$$\bar{p} = \{p(i) \mid i = 1, \dots, N_{pop}\} \quad (3)$$

where N_{pop} is the population size.

Starting from an initial population \bar{p}^0 , the SGA iteratively constructs populations $\bar{p}^k, k=1..N_{gen}$, with N_{gen} denoting the total number of SGA generations. Subsequent generations are constructed by iteratively acting upon \bar{p}^0 with a set of genetic operators. The operators that induce the transition $\bar{p}^k \rightarrow \bar{p}^{k+1}$ are guided solely by knowledge of the vector of objective function values

$$f^k = \left\{ f\left(x\left(p^k(i)\right)\right) \mid i = 1..N_{pop} \right\} \quad (4)$$

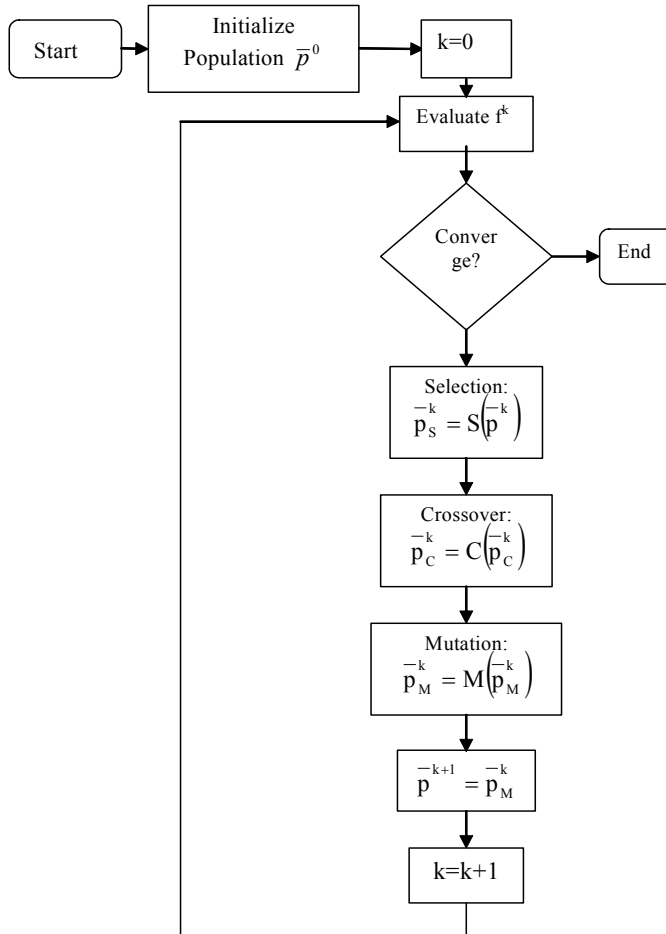


Fig. 9. The flowchart of a simple GA

and induce changes in the genetic makeup of the population leading to a \bar{p}^{k+1} comprising individuals that are, on average better adapted to their environment than those in \bar{p}^k , i.e., they are characterized by higher objective function values.

This change is effected by three operators mentioned in the introduction: selection (S), crossover (C), and mutation (M).

The selection operator implements the principle of survival of the fittest. Acting on \bar{p}^k , S produces a new population $\bar{p}_S^k = S(\bar{p}^k)$ again of size N_{pop} that is, on average, populated by the better-fit individuals present in \bar{p}^k . Among the many existing schemes tournament selection has been chosen. The crossover operator mimics natural procreation. Specifically, C acts upon the population \bar{p}_S^k by mating its members, thereby creating a new population

$$\bar{p}_C^k = \bigcup_{i=1}^{N_{pop}/2} C\left(ch\left(\bar{p}_S^k\right), ch\left(\bar{p}_S^k\right)\right) \quad (5)$$

where the chromosome crossover operator C selects a random crossover allele $a_{N_{cross}}$ between the two chromosomes to be crossed upon which it acts with probability P_{cross} . The mutation operator generates a new population of size by introducing small random changes into \bar{p}_C^k . The action of M can be represented in operator form as

$$\bar{p}_M^{-k} = \bigcup_{i=1}^{N_{pop}} M \left(\bar{p}_C^{-k}(i) \right) \quad (6)$$

The cost function of the optimization procedure has been the coverage percentage of the points for which the received power is greater than a given level.

$$c(P_{rec}) = \frac{\text{Number of points } (P_{thresh.} < P_{rec})}{\text{Total number of test points}} \quad (7)$$

The number of test points to evaluate the cost function above was 12000 on the floor level, and the P_{thresh} level was -70 dBm, respectively.

4.2 DIRECT algorithm

The DIRECT optimization algorithm is a derivative-free global algorithm that yields a deterministic and unique solution (Daniel E. Finkel, 2003). Its attribute of possessing both local and global properties make it ideal for fast convergence. An essential aspect of the DIRECT algorithm is the subdivision of the entire design space into hyper-rectangles or hyper-cubes for multidimensional problems.

The iteration starts by choosing the center of the design space as the starting point. Subsequently, at each iteration step, DIRECT selects and subdivides the set of hyper-cubes that are most likely to produce the lowest objective function. This estimation is based on Lipschitzian optimization method. Basically for one dimension a function is called Lipschitz continuous on domain R with Lipschitz constant α if

$$|f(x_1) - f(x_2)| \leq \alpha |x_1 - x_2| \quad x_1, x_2 \in R \quad (8)$$

where

$f(x)$ is the objective function for the optimization problem.

The complementary of the coverage percentage which has to be minimized was chosen as objective function for the DIRECT algorithm.

$$f(x) = 1 - c(P_{rec}) \quad (9)$$

The Lipschitzian function finds the global minimum point provided the constant α is specified to be greater than the largest rate of change of the objective function within the design space and that the objective function value is continuous. Within DIRECT, all possible values of the Lipschitzian constant α are used with the larger values of α chosen for global optimization (to find the basin of convergence of the optimum) followed by smaller values of α for local optimizations within this basin of convergence. As mentioned above, DIRECT divides the domain into multiple rectangles at each iteration. Thus, the convergence process is greatly sped up and the optimization algorithm achieves both local and global searching properties.

As illustration of subdividing the search region into hyper-rectangles and sampling, two dimensional problem optimization steps are shown in Fig. 10.

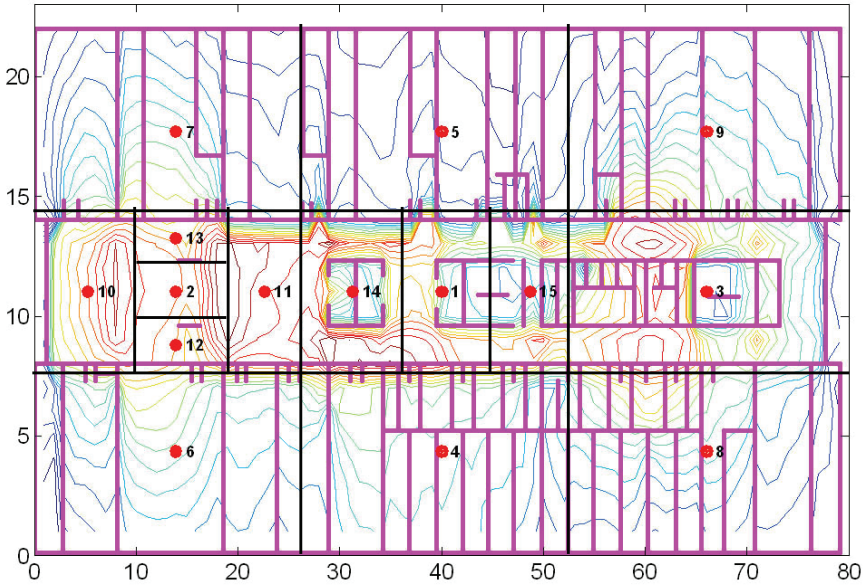


Fig. 10. DIRECT global optimizer search steps

The Algorithm DIRECT is stated as follows.

```

Algorithm DIRECT
Start point at the center of the user defined area ( $0 \leq x \leq 80; 0 \leq y \leq 22$ )
while  $f_{\text{objective}} > f_{\text{limit}}$  and  $\text{iteration steps} < \text{iteration steps}_{\text{limit}}$ 
    Divide the area of investigation space into three rectangles
    Set the centers of the three rectangles
    Use the Lipschitz constant  $\alpha$  to select the rectangle has to be divided
end while
    
```

4.3 Hierarchic two level optimization method

The new proposed hierarchic optimization method uses a two level optimization procedure, in which the radio wave propagation models differ. The propagation models are:

- indoor power law,
- Motley-Keenan, (Motley & Keenan 1990)

The optimization procedure is based on simple GA and starts with simple power law model. At the point of non changing cost function the procedure switches to the more sophisticated model to Motley-Keenan model. In line with model change the mutation and crossover probabilities are decreased also.

A single approach of ITU-R model is used (ITU, 1238), except that the propagation exponent is accounted for explicitly by changing the path loss exponent. The model is assumed to produce the following total path loss model (in decibels)

$$L = 20\log\left(f_c^{[MHz]}\right) + 30\log\left(r^{[m]}\right) - 28 \quad (10)$$

where r is the distance between transmitter and receiver antennas.

The indoor power law model takes only account the distance to predict the received power therefore the first level of optimization procedure ends a coverage picture as can be seen on Fig. 11. for five AP-s. In this way the expensive propagation model take place only at refinement of the AP positions.

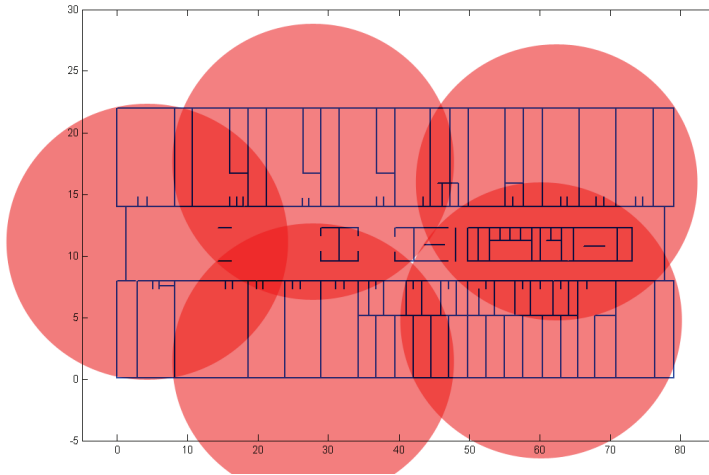


Fig. 11. Access Points positions after the first level optimization

The second optimization step assigns the AP regions based on the previous AP positions, as can be seen on Fig. 12.

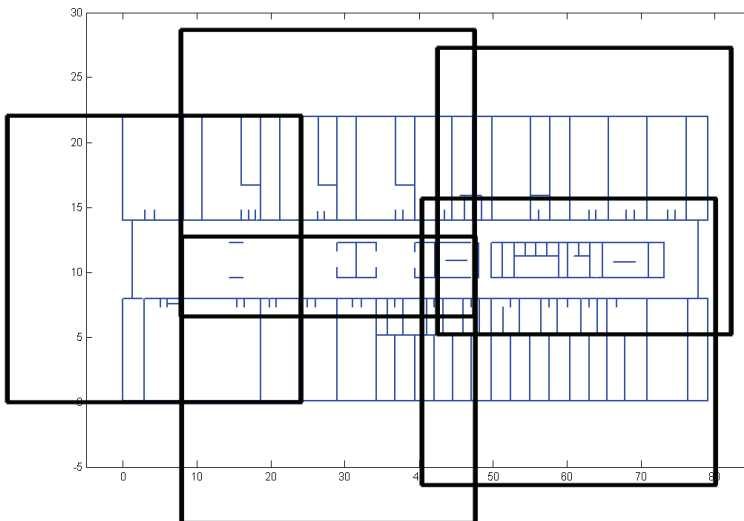


Fig. 12. AP regions for second level optimization

The second optimization step uses the Motley-Keenan model, which regression parameters have been determined using Ray Launching deterministic radiowave propagation model. The most important innovation of the two steps method is the decrease the search area in the first search period using a homogeneous wave propagation model and to get quick results on candidate areas for access points for the second phase of search which is using genetic algorithm as well.

5. Results

The testing of the SGA optimization has been done with two testing cases at the office building in which first optimizing the coverage for part of the floor area and secondly for the whole level.

The results are shown for population size of 14, crossover probability - 0.12, mutation probability - 0.01, simple roulette wheel selection and simple elitist strategy.

The first scenario is an optimization on AP positions (circles in Fig. 13.) of the half part of the floor. The Fig. 13 shows the original 4 AP positions which were chosen to best coverage in laboratories and the corridor coverage was not an aim. The Fig. 14. shows the optimal AP positions using the cost function of (Eq. 7). The simulated distribution of received power for the two geometries is shown in Fig. 15-16. with the measured results.

To make the measurements we have chosen WLAN APs and the power levels were measured using laptops with external wireless adapter moved on the area of investigation. 90 sampling points in distances of 1 m were chosen on the level and the comparison of Fig. 15. and 16. show a good agreement for the received power distribution.

The most important change in the distributions of optimized and not optimized cases is increased number of points with proper coverage. (Table 2.)

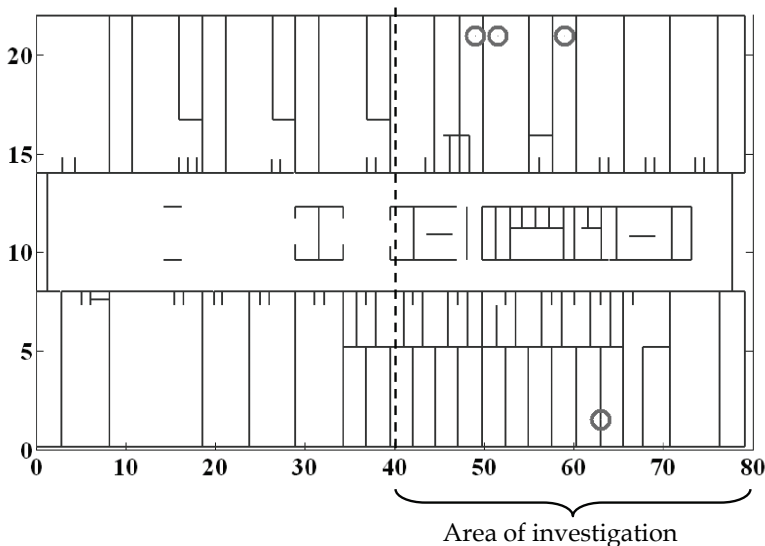


Fig. 13. Original (not optimized) AP positions

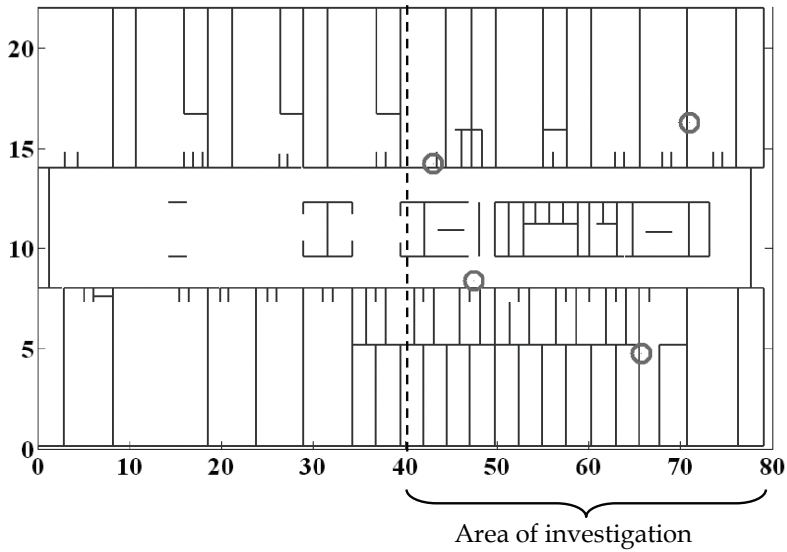


Fig. 14. Optimized AP positions

Configuration	Not optimized	Optimized
Coverage for $P_{rec} > -60\text{dBm}$ (simulation)	40%	75%
Coverage for $P_{rec} > -60\text{dBm}$ (measurement)	50%	80%

Table 2. Area Coverage for Optimized and not Optimized Case

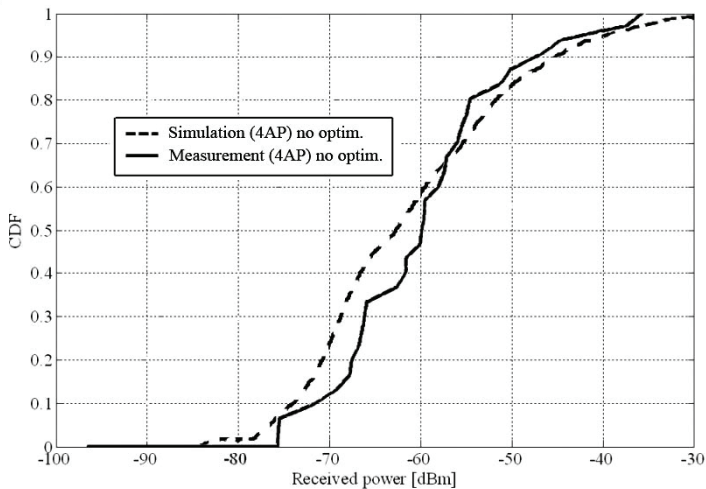


Fig. 15. Cumulative Density Function of received power level (not optimized)

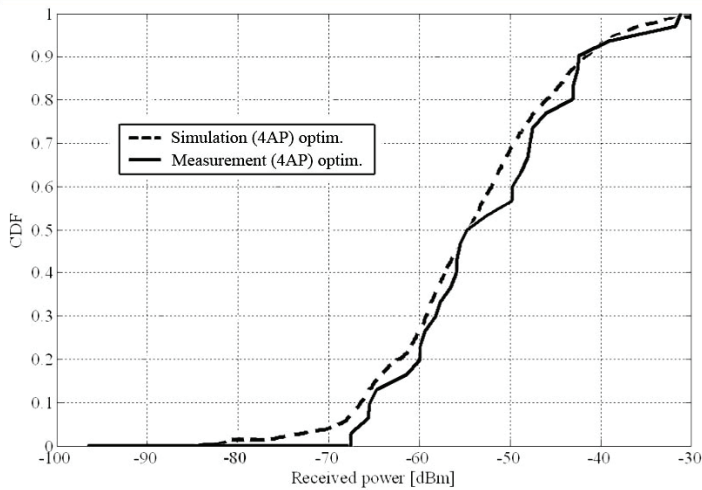


Fig. 16. Cumulative Density Function of received power level (optimized)

The convergence of the Genetic Algorithm can be improved by adjusting the crossover and mutation probability. The Fig. 17. shows the convergence dependence on these parameters for the same generation size.

The Fig. 17. shows a significant dependence of convergence on GA parameters and this results in a 1 to 10 running time ratio. The iteration step means that the number of necessary objective function evaluation can be calculated by multiplying with the population size.

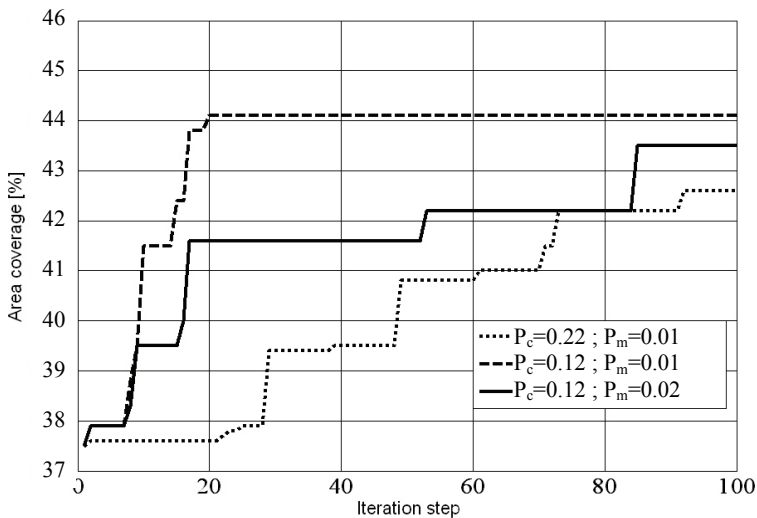


Fig. 17. Genetic Algorithm convergence

The second simulation is on the entire floor level and the aim of the simulation is to compare the necessary number of APs for the same area coverage.

The Fig. 18. shows plausible positions of APs and the Fig. 19. the optimized ones.

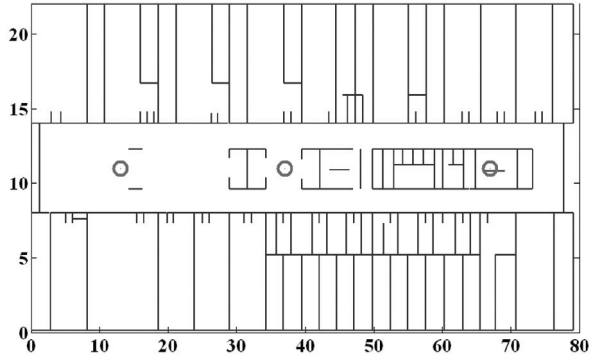


Fig. 18. Plausible AP positions

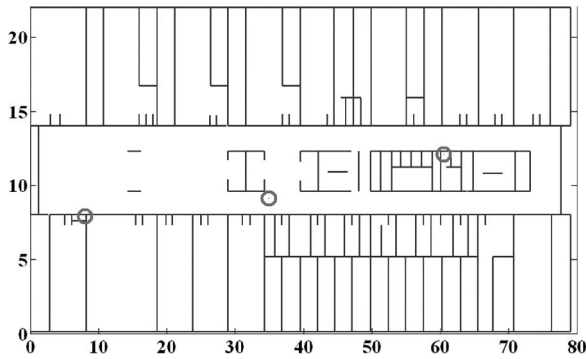


Fig. 19. Optimized AP positions

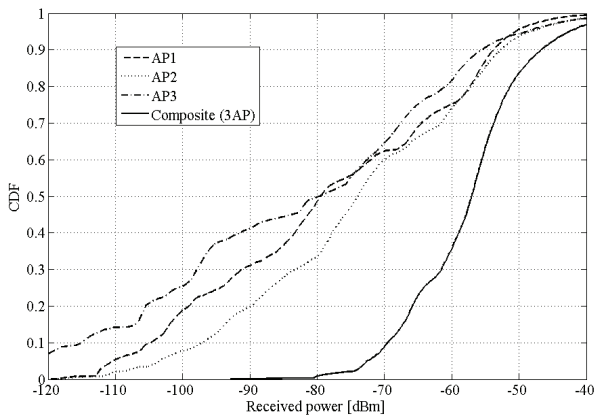


Fig. 20. Independent and composite CDF (optimized AP positions)

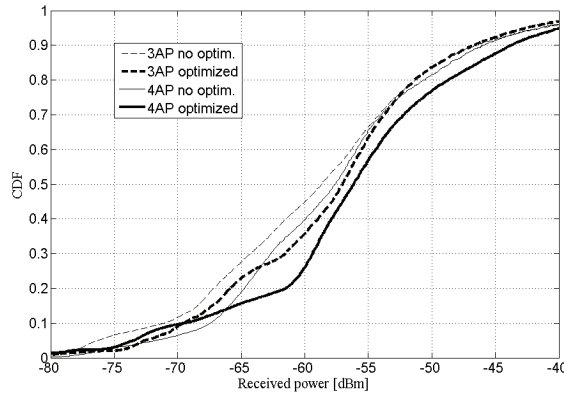


Fig. 21. Optimized and not optimized CDF using 3 and 4 APs

The Fig. 21. and Table 3 summarizes the importance of RU position of HFR. With the proper choice of the placement the optimized 3 AP network configuration results nearly the same coverage as the configuration 6 AP with APs installed in plausible positions.

Configuration	3AP	4AP	6AP
Coverage (not optimized)	55%	60%	66%
Coverage (optimized)	65%	75%	87%

Table 3. Area coverage for optimized and not optimized cases

These results (Table 3.) illustrate and justify well the importance of Remote Unit or Access Point installation positions in HFR networks in order to maximize the wireless coverage. Using the mentioned optimization procedure the network cost can be significantly reduced and the optical distribution network also can be simplified.

As we have shown the SGA is a powerful global optimization tool to improve the indoor coverage for HFR and other mobile radio systems. The main drawback of the method is the ambiguous convergence and therefore its application needs experience of the user. The DIRECT global optimization algorithm is a derivative-free global algorithm that yields a deterministic and unique solution. In the next simulation results will be shown using DIRECT for the same indoor AP position optimization problem. We are comparing DIRECT to SGA and the main point of comparison is the number of evaluation of objective function. It is worth to investigate the candidate points for the AP position by the DIRECT algorithm. The simplest case is analyzed for one AP network and the investigated and best candidate points are shown in accordance with the objective function the area of coverage % in Fig. 24. The objective function was only evaluated 1 by 1 m resolution. It is well appreciable the testing of the attractive AP positions with high area coverage property.

Next the convergence of SGA and DIRECT will be compared in Fig. 25, 26 and 27. It can be point out that the DIRECT algorithm behaves well for 1 or 2 AP optimization problems (i.e. for 2 and 4 dimensional optimizations) but the convergence rate achieve is far below the SGA for 3 AP problem. Similar behavior can be experienced for higher dimensional optimization problems. Based on this investigations DIRECT algorithm can be proposed for

low dimensional cases till 4 dimensions but the theoretically guaranteed fast and unique solution of global problem has to analyzed further.

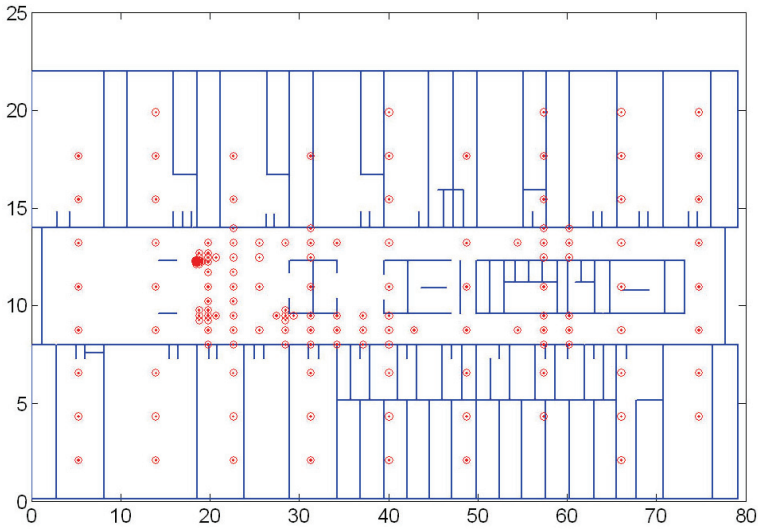


Fig. 22. Candidate points for AP position (after 12, 24, 36...iterations, DIRECT)

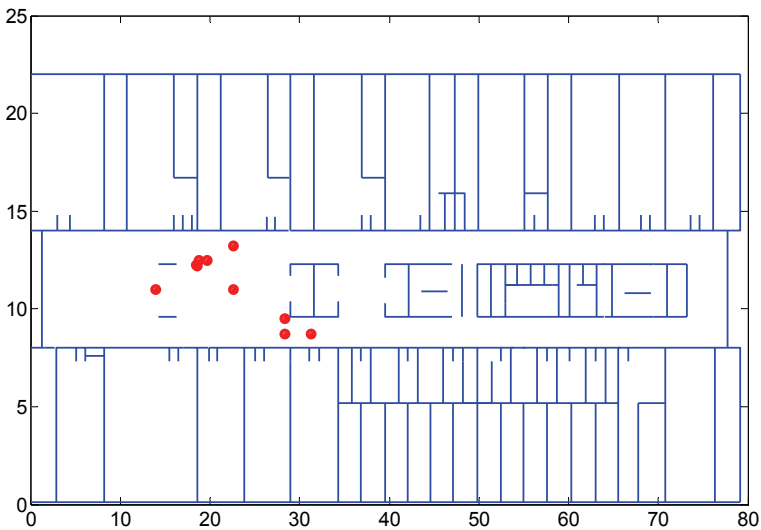


Fig. 23. Best candidate point for AP position (after 12, 24, 36...iterations, DIRECT)

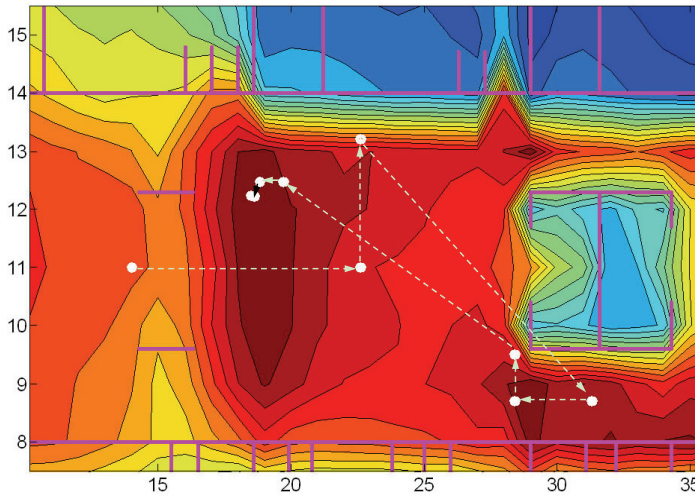


Fig. 24. Best candidate points for AP position (Area of coverage % is also shown)

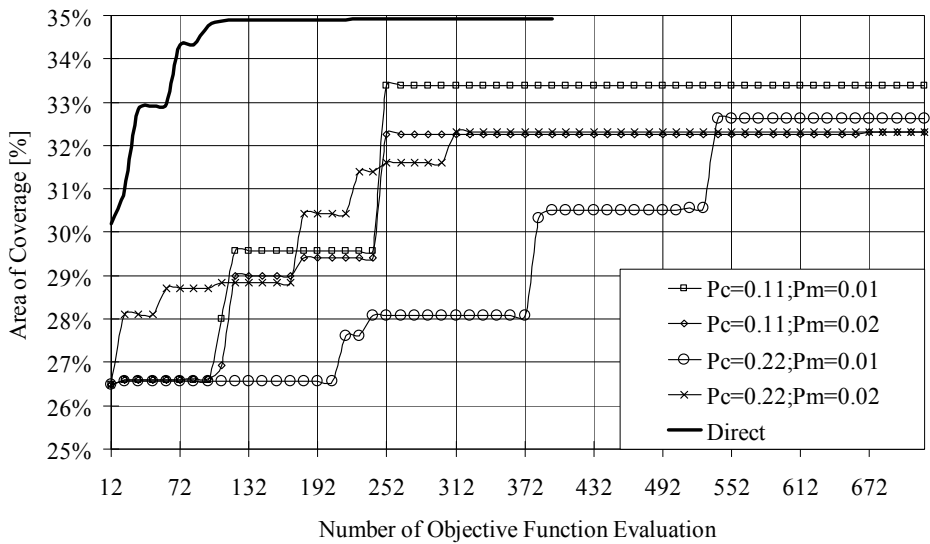


Fig. 25. Convergence of SGA and DIRECT for 1 Access Point

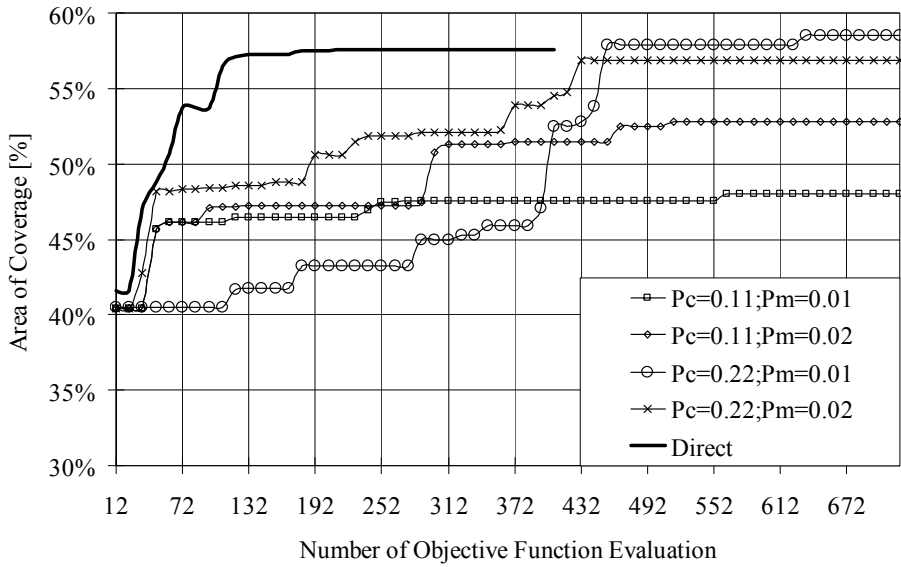


Fig. 26. Convergence of SGA and DIRECT for 2 Access Points

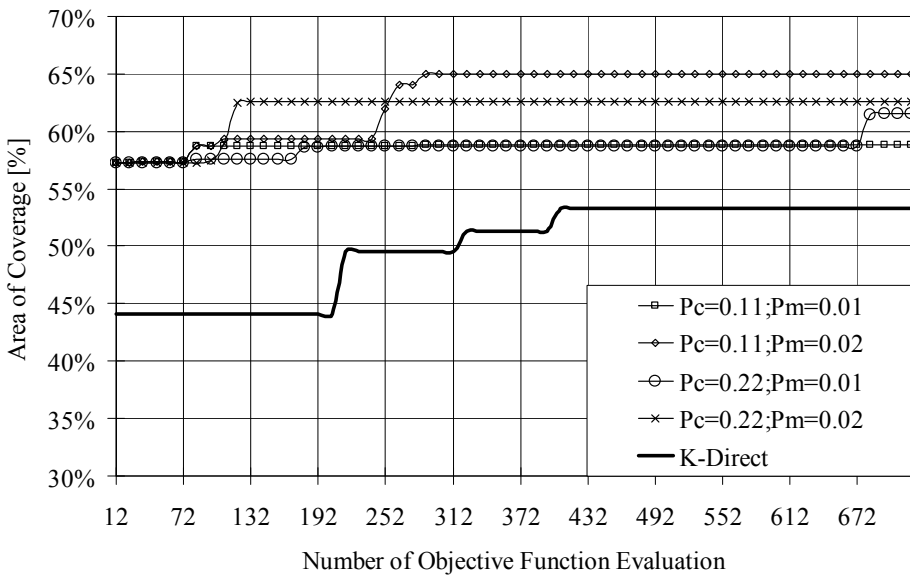


Fig. 27. Convergence of SGA and DIRECT for 3 Access Points

The next part shows the comparisons of SGA and the proposed hierarchic two steps optimization method, first the convergence of the simple Genetic Algorithm for different population sizes (Fig. 28.). Now we investigate 6 AP optimization cases in order to validate the two step method. As we have seen problems of dimensions above 4 can not be analyzed with DIRECT and therefore for comparison this 12 dimensional problem will be investigated. First the GA optimization is shown after performing the AP search by using power law path loss model i.e. the hierarchic approach. Finally the optimization results are analyzed.

Fig. 28 presents effect of values population size, crossover (C) and mutation (M) probability on convergence for 6 APs placement single GA optimization in our simulations.

The population size extension effect a better convergence (Figure 28) but the calculation time increase polynomial.

The most important observations are that the crossover and mutation probabilities have optimal values in this case for the geometry investigated these values are $P_C=0.22$ and $P_M=0.01$. (Fig. 28-29)

Next the proposed hierarchic optimization model is shown after performing the AP search by using power law path loss model. The limitation of the optimization areas for the AP positions results in much better convergence for each configuration (Fig. 30-31).

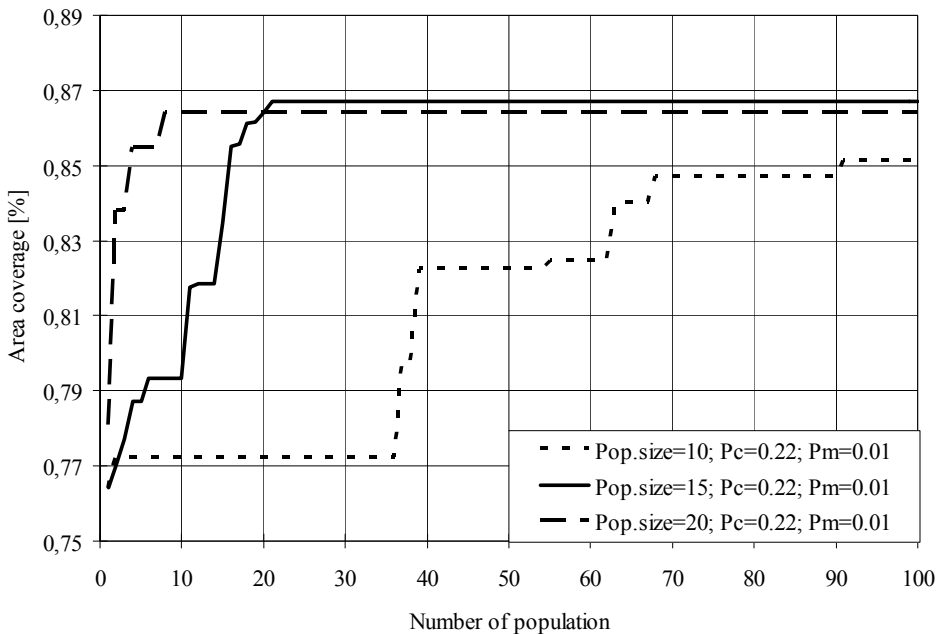


Fig. 28. Genetic Algorithm convergence (6AP whole floor)

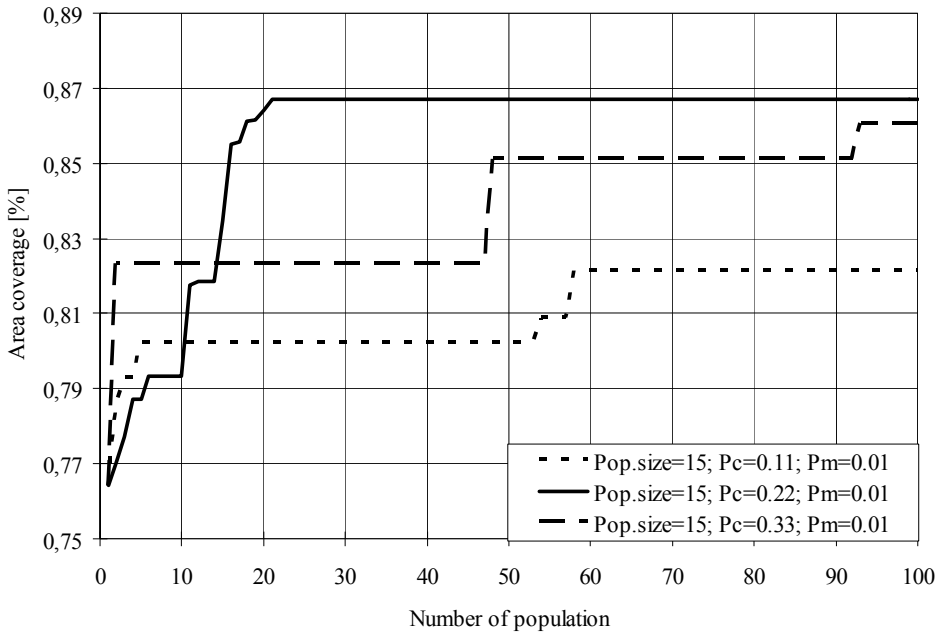


Fig. 29. Genetic Algorithm convergence (6AP whole floor)

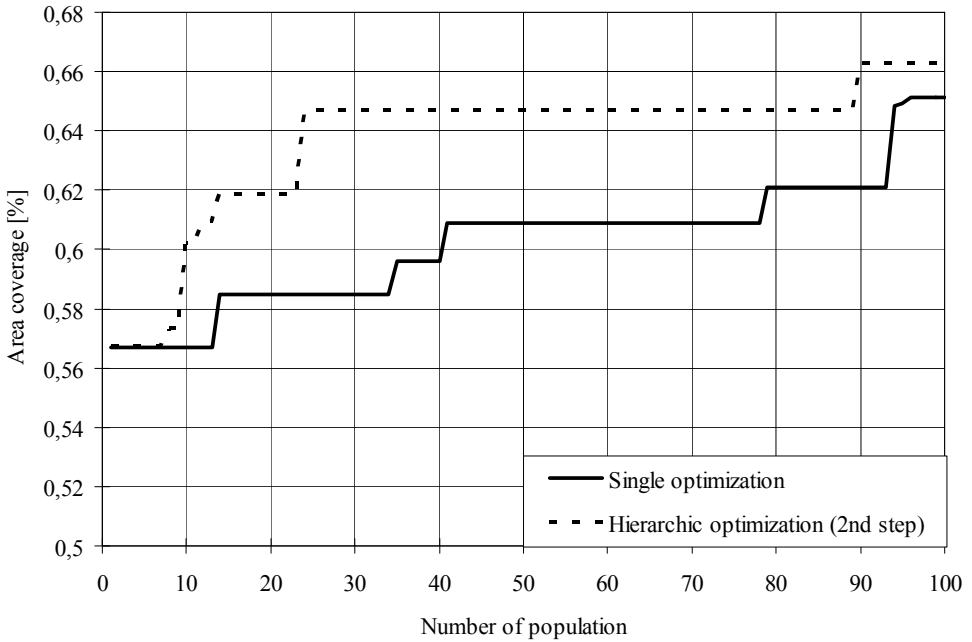


Fig. 30. Genetic Algorithm convergence (3AP whole floor)

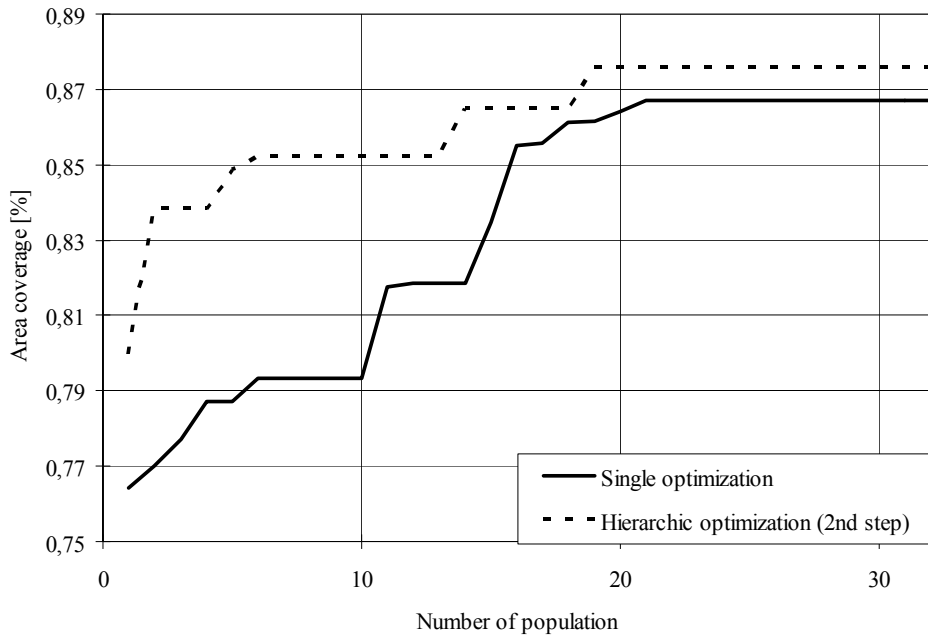


Fig. 31. Genetic Algorithm convergence (6AP whole floor)

6. Conclusion

The optimal Remote Unit position of Hybrid Fiber Radio is investigated for indoor environment. The article illustrates the possibility of optimization of HFR network using Genetic Algorithm in order to determine positions of APs. Two new approaches are introduced to solve the global optimization problem the DIRECT and a hierarchic two step optimization combined with genetic algorithm. The methods are introduced and investigated for 1,2, 3 and 6 AP cases. The influence of Genetic Algorithm parameters on the convergence has been tested and the optimal radio network is investigated. It has been shown that for finding proper placement the necessary number of APs can be reduced and therefore saving installation cost of WLAN or HFR.

It has been shown that for finding proper placement the necessary number of RU can be reduced and therefore saving installation cost of HFR. The results clearly justify the advantage of the method we used but further investigations are necessary to combine and to model other wireless network elements like leaky cables, fiber losses. Other promising direction is the extension of the optimization cost function with interference parameters of the wireless network part and with outer interference.

7. References

- Martin D. Adickes, Richard E. Billo, Bryan A. Norman, Sujata Banerjee, Bartholomew O. Nnaji, Jayant Rajgopal (2002). Optimization of indoor wireless communication network layouts, IIE Transactions, Volume 34, Number 9 / September, 2002, Springer,
- Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009-2014, (white paper), 2010
http://cisco.biz/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf
- Lóránt Farkas, István Laki, Lajos Nagy (2001). Base Station Position Optimization in Microcells using Genetic Algorithms, ICT'2001, 2001, Bucharest, Romania
- Daniel E. Finkel (2003). DIRECT Optimization Algorithm User Guide,
<http://www.ncsu.edu/crsc/reports/ftp/pdf/crsc-tr03-11.pdf>
- J.M. Keenan, A.J. Motley (1990). Radio Coverage in buildings, BT Tech. J., 8(1), 1990, pp. 19-24.
- Z. Michalewicz (1996). Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, Berlin, 1996.
- E. Michielssen, Y. Rahmat-Samii, D.S. Weile (1999). Electromagnetic System Design using Genetic Algorithms, Modern Radio Science, 1999.
- R.D. Murch, K.W. Cheung (1996). Optimizing Indoor Base-station Locations, XXVth General Assembly of URSI, 1996, Lille, France
- Lajos Nagy, Lóránt Farkas (2000). Indoor Base Station Location Optimization using Genetic Algorithms, PIMRC'2000 Proceedings, Sept. 2000, London, UK
- A. Portilla-Figueroa, S. Salcedo-Sanz, Klaus D. Hackbarth, F. López-Ferreras, and G. Esteve-Asensio (2009). Novel Heuristics for Cell Radius Determination in WCDMA Systems and Their Application to Strategic Planning Studies, EURASIP Journal on Wireless Communications and Networking, Volume 2009 (2009)

- Liza K. Pujji, Kevin W. Sowerby, Michael J. Neve (2009). A New Algorithm for Efficient Optimization of Base Station Placement in Indoor Wireless Communication Systems, 2009 Seventh Annual Communication Networks and Services Research Conference, Moncton, New Brunswick, Canada, ISBN: 978-0-7695-3649-1
- R. E. Schuh, D. Wake, B. Verri and M. Mateescu, Hybrid Fibre Radio Access (1999) A Network Operators Approach and Requirements, 10th Microcoll Conference, Microcoll'99, Budapest, Hungary, pp. 211-214, 21-24 March, 1999
- Yufei Wu, Samuel Pierre (2007). Optimization of 3G Radio Network Planning Using Tabu Search, Journal of Communication and Information Systems, Vol. 22, No. 1, 2007

Introduction to Packet Scheduling Algorithms for Communication Networks

Tsung-Yu Tsai¹, Yao-Liang Chung² and Zsehong Tsai²

¹*Institute for Information Industry*

²*Graduate Institute of Communication Engineering, National Taiwan University*

^{1,2}*Taipei, Taiwan, R.O.C.*

1. Introduction

As implied by the word “packet scheduling”, the shared transmission resource should be intentionally assigned to some users at a given time. The process of assigning users’ packets to appropriate shared resource to achieve some performance guarantee is so-called packet scheduling.

It is anticipated that packetized transmissions over links via proper packet scheduling algorithms will possibly make higher resource utilization through statistical multiplexing of packets compared to conventional circuit-based communications. A packet-switched and integrated service environment is therefore prevalent in most practical systems nowadays. However, it will possibly lead to crucial problems when multiple packets associated to different kinds of Quality of Service (QoS) (e.g. required throughput, tolerated delay, jitter, etc) or packet lengths competing for the finite common transmission resource. That is, when the traffic load is relatively heavy, the first-come-first-serve discipline may no longer be an efficient way to utilize the available transmission resource to satisfy the QoS requirements of each user. In such case, appropriate packet-level scheduling algorithms, which are designed to schedule the order of packet transmission under the consideration of different QoS requirements of individual users or other criteria, such as fairness, can alter the service performance and increase the system capacity. As a result, packet scheduling algorithms have been one of the most crucial functions in many practical wired and wireless communication network systems. In this chapter, we will focus on such topic direction for complete investigation.

Till now, many packet scheduling algorithms for wired and wireless communication network systems have been successfully presented. Generally speaking, in the most parts of researches, the main goal of packet scheduling algorithms is to maximize the system capacity while satisfying the QoS of users and achieving certain level of fairness. To be more specific, most of packet scheduling algorithm proposed are intended to achieve the following desired properties:

1. Efficiency:

The basic function of packet scheduling algorithms is scheduling the transmission order of packets queued in the system based on the available shared resource in a way that satisfies the set of QoS requirements of each user. A packet scheduling algorithm is generally said to

be more efficient than others if it can provide larger capacity region. That is, it can meet the same QoS guarantee under a heavier traffic load or more served users.

2. Protection:

Besides the guarantees of QoS, another desired property of a packet scheduling algorithm to treat the flows like providing individual virtual channels, such that the traffic characteristic of one flow will have as small effect to the service quality of other flows as possible. This property is sometimes referred as *flow isolation* in many scheduling contexts. Here, we simply define the term flow to be a data connection of certain user. A more formal definition will be given in the next section.

Flow isolation can greatly facilitate the system to provide flow-by-flow QoS guarantees which are independent of the traffic demand of other flows. It is beneficial in several aspects, such as the per-flow QoS guarantee can be avoided to be degraded by some ill-behavior users which send packet with a higher rate than they declared. On the other hand, a more flexible performance guarantee service scheme can also be allowed by logically dividing the users which are associated to a wide range of QoS requirements and traffic characteristic while providing protection from affecting each other.

3. Flexibility:

A packet scheduling algorithm shall be able to support users with widely different QoS requirements. Providing applications with vast diversity of traffic characteristic and performance requirements is a typical case in most practical integrated systems nowadays.

4. Low complexity:

A packet scheduling algorithm should have reasonable computational complexity to be implemented. Due to the fast growing of bandwidth and transmission rate in today's communication system, the processing speed of packets becomes more and more critical. Thus, the complexity of the packet scheduling algorithm is also of important concern.

Due to the evolution process of the communication technology, many packet scheduling algorithms for wireless systems in literatures are based on the rich results from the packet scheduling algorithms for wired systems, either in the design philosophy or the mathematical models. However, because of the fundamental differences of the physical characteristics and transmission technologies used between wired and wireless channels, it also leads to some difference between the considerations of the packet scheduling for wired and wireless communication systems. Hence, we suggest separate the existing packet scheduling algorithms into two parts, namely, wired ones and wireless ones, and illustrate the packet scheduling algorithms for wired systems first to build several basic backgrounds first and then go to that for the wireless systems.

The rest of the chapter is outlined as follows. In Section 2, we will start by introducing some preliminary definition for preparation. Section 3 will make an overview for packet scheduling algorithms in wired communication systems. Comprehensive surveys for packet scheduling in wireless communication systems will then be included in Section 4. In Section 5, we will employ two case studies for designing packet scheduling mechanisms in OFDMA-based systems. In Section 6, summary and some open issues of interest for packet scheduling will be addressed. Finally, references will be provided in the end of this chapter.

2. Preliminary definitions

The review of the packet scheduling algorithms throughout this chapter considers a packet-switched single server. The server has an outgoing link with transmission rate C . The main

task of the server is dealing with the packets input to it and forwarding them into the outgoing link. A packet scheduling algorithm is employed by the server to schedule the appropriate forwarding order to the outgoing link to meet a variety of QoS requirements associated to each packet. For wireline systems, the physical medium is in general regarded as stable and robust. Thus the packet error rate (PER) is usually ignored and C can be simply considered as a constant with unit bits/sec. This kind of model is usually referred as *error-free channel* in literatures. On the other hands, for wireless systems, the situation can become much more complicate. Whether in wireless networks with short transmission range (about tens of meters) such as WLAN and femtocell or that with long transmission range (about hundreds of meters or even several kilometers) such as the macrocell environments based on WCDMA, WiMAX and LTE, the packet transmission in wireless medium suffers location-dependent path loss, shadowing, and fading. These impairment make the PER be no longer ignorable and the link capacity C may also become varying (when adaptive modulation and coding is adopted). This kind of model is usually referred as *error-prone channel* in literatures.

Each input packet is associated to a *flow*. Flow is a *logical* unit which represents a sequence of input packets. In practice, packets associated to the same flows often share the same or similar quality of service (QoS) requirement. There should be a *classifier* in the server to map each input packets to appropriate flows.

The QoS requirement of a flow is usually characterized by a set of *QoS parameters*. In practice, the QoS parameters may include tolerant delay or tolerant jitter of each packet, or data rate requirement such as the minimum required throughput. The choice of QoS parameters might defer flow by flow, according to the specific requirement of different services. For example, in IEEE 802.16e [47], each data connection is associated to a service type. There are totally five service types to be defined. That is, unsolicited grant service (UGS), real-time polling service (rtPS), extended real-time polling service (ertPS), non-real-time polling service (nrtPS), and best effort (BE). Among these, rtPS is generally for streaming audio or video services, and the QoS parameters contains the minimum reserved rate, maximum sustained rate, and maximum latency tolerant. On the other hands, UGS is designed for IP telephony services without silence suppression (i.e. voice services with constant bit rate). The QoS parameters of UGS connections contains all the parameters of rtPS connections and additionally, it also contains a parameter, jitter tolerance, since the service experiment of IP telephony is more sensitive to the smoothness of traffic. Moreover, for nrtPS, which is mainly designed for non-real-time data transmission service such as FTP, the QoS parameters contains minimum reserved data rate and maximum sustained data rate. Unlike rtPS and UGS, which required the latency of each packet to be below certain level, nrtPS is somewhat less sensitive to the packet latency. It allows some packets to be postponed without degrading the service experiment immediately, however, an average data rate should still be guaranteed, since throughput is of the most concern for data transmission services.

The server can be further divided into two categories, according to the *eligible time* of the input packets. Eligible time of a packet is defined as the earliest time that the packet begins being transmitted. Additionally, a packet is called eligible when it is available to be transmitted by the server. If all packets immediately become eligible for transmission upon arrival, the system is called *work-conserving*, otherwise, it is called *nonwork-conserving*. A direct consequence of a system being work-conserving is that the server is never idle whenever there are packets queued in the server. It always forwards the packets when the queues are not empty.

3. Packet scheduling algorithms in wireline systems

In this section, we will introduce several representative packet scheduling algorithms of wireline systems. Their merits and expense will be examined respectively.

3.1 First Come First Serve (FCFS)

FCFS may be the simplest way for a scheduler to schedule the packets. In fact, FCFS does not consider the QoS parameters of each packets, it just sends the packets according to the order of their arrival time. Thus, the QoS guarantee provided by FCFS is in general weak and highly depends on the traffic characteristic of flows. For example, if there are some flows which have very bursty traffic, under the discipline of FCFS, a packet will very likely be blocked for a long time by packets burst which arrives before it. In the worst case, the unfairness between different flows cannot be bounded, and the QoS cannot be no longer guaranteed. However, since FCFS has the advantage of simple to implement, it is still adopted in many communication networks, especially the networks providing *best effort* services. If some level of QoS is required, then more sophisticated scheduling algorithm is needed.

3.2 Round Robin

Round Robin (RR) scheme is a choice to compensate the drawbacks of FCFS which also has low implementation complexity. Specifically speaking, newly arrival packets queue up by flow such that each flow has its respective queue. The scheduler polls each flow queue in a cyclic order and serves a packet from any-empty buffer encountered; therefore, the RR scheme is also called flow-based RR scheme. RR scheduling is one of the oldest, simplest, fairest and most widely used scheduling algorithms, designed especially for time-sharing systems. They do offer greater fairness and better bandwidth utilization, and are of great interest when considering other scenarios than the high-speed point-to-point scenario. However, since RR is an attempt to treat all flows equally, it will lead to the lack of flexibility which is essential if certain flows are supported to be treated better than other ones.

3.3 Strict priority

Strict priority is another classical service discipline which assigns *classes* to each flow. Different classes may be associated to different QoS level and have different *priority*. The eligible packets associated to the flow with higher-priority classes are sent ahead of the eligible packets associated to the flow with lower-priority classes. The sending order of packets under strict priority discipline only depends on the classes of the packets. This is why it called "strict" since the eligible packets with lower-priority classes will never be sent before the eligible packets with higher-priority classes. Strict priority suffers from the same problem as that of FCFS, since a packet may also wait arbitrarily long time to be sent. Especially for the packets with lower-priority classes, they may be even starved by the packets with higher-priority classes.

3.4 Earliest Deadline First (EDF)

For networks providing real-time services such as multimedia applications, earliest deadline first (EDF) [5][6] is one of the most well-known scheduling algorithms. Under EDF discipline, each flow is assigned a tolerant delay bound d_i ; a packet j of flow i arriving at time a_{ij} is naturally assigned a deadline $a_{ij} + d_i$. Each eligible packet is sent according to the

increasing order of their deadlines. The concept behind EDF is straightforward. It essentially schedules the packets in a greedy manner which always picks the packets with the closest deadline. Compare with strict priority discipline, we can regard EDF as a scheduling algorithm which provides *time-dependent priority* [8] to each eligible packet. Actually, the priority of an eligible packet under EDF is an increasing function of time since the sending order in EDF is according to the closeness of packets' deadlines. This fact allows the guarantee of QoS if the traffic characteristic of each flow obeys some specific constraint (e.g. the incoming traffic in a time interval is upper bounded by some amount). Define the traffic envelope $A_i(t)$ is the amount of flow i traffic entering the server in any interval of length t . The authors in [9] and [13] proved that in a work-conserving system, the necessary and sufficient condition for the served flows are schedulable (i.e. each packet are guaranteed to be sent before its deadline expires), which is expressed by

$$\sum_i A_i(t - d_i) + l_{\max} I_{\{d_{\min} \leq t \leq d_{\max}\}} \leq Ct \quad (3.1)$$

where C is the outgoing link capacity as described in section 2, l_{\max} is the maximum possible packet size among all flows, $d_{\min} = \min_i\{d_i\}$, $d_{\max} = \max_i\{d_i\}$, $I_{\{event\}}$ is the indicator function of event E .

An important result of EDF is that it has been known to be the optimal scheduling policy in the sense that it has the largest *schedulable region* [9]. More specifically, given N flows with traffic envelopes $A_i(t)$ ($i = 1, 2, \dots, N$), and given a vector of delay bounds $\mathbf{d} = (d_1, d_2, \dots, d_N)$, where d_i is the to delay bound that flow i can tolerate. It can be proved that if \mathbf{d} is schedulable under a scheduling algorithm π , then \mathbf{d} will also be schedulable under EDF.

Although EDF has optimal schedulable region, it encounters the same drawback as that of FCFS and strict priority disciplines. That is, the lack of protection between flows which introduces weak flow isolation (see section 1). For example, if some flows do not have bounded traffic envelope, that is, $A_i(t)$ can be arbitrary large (or at least, very large) for some i , then the condition in (3.1) can't no longer be guaranteed to be satisfied, and no QoS guarantee can be provided to any flows being served. In the next section, we will introduce generalized processor sharing (GPS) discipline, which can provide ideal flow isolation property. The lack of flow isolation of EDF is often compensated by adopting *traffic shapers* to each flow to shape the traffic envelopes and bound the worst-case amount of incoming traffic of per flow. There are also some modified versions of EDF proposed to provide more protection among flows, such as [7] [10].

3.5 Generalized Processor Sharing (GPS)

Generalized processor sharing (GPS) is an ideal service discipline which provides perfect flow isolation. It assumes that the traffic is infinitely divisible, and the server can serve multiple flows simultaneously with rates proportional to the *weighting factors* associated to each flow. More formally, assume there are N flows, and each flow i is characterized by a weighting factor w_i . Let $S_i(\tau, t)$ be the amount of flow i traffic served in an interval (τ, t) and a flow is *backlogged* at time t if a positive amount of that flow's traffic is queued at time t . Then, a GPS server is defined as one service discipline for which

$$\frac{S_i(\tau, t)}{S_j(\tau, t)} \geq \frac{w_i}{w_j}, j = 1, 2, \dots, N \quad (3.2)$$

For any flow i that is continuously backlogged in the interval (τ, t) .
Summing over all flow j , we can obtain:

$$S_i(\tau, t) \sum_j w_j \geq (t - \tau) C w_i$$

that is, when flow i is backlogged, it is guaranteed a minimum rate of

$$g_i = \frac{w_i}{\sum_j w_j} C$$

In fact, GPS is more like an idealized model rather than a scheduling algorithm, since it assumes a fluid traffic model in which all the packets is infinitely divisible. The assumptions make GPS not practical to be realized in a packet-switched system. However, GPS is still worth to remark for the following reasons:

1. It provides following attractive ideal properties and can be a benchmark for other scheduling algorithms.

a. Ideal resource division and service rate guarantee

GPS assumes that a server can serve all backlogged flows simultaneously and the outgoing link capacity C can be perfectly divided according to the weight factor associated to each backlogged flow. It leads to ideal flow isolation in which each flow can be guaranteed a minimum service rate independent of the demands of the other flows. Thus, the delay of an arriving bit of a flow can be bounded as a function of the flow's queue length, which is independent of the queue lengths and arrivals of the other flows. According to this fact, one can see that if the traffic envelope of a flow obeys some constraint (e.g. leaky buckets) and is bounded, then the traffic delay of a flow can be guaranteed. Schemes such as FCFS and strict priority do not have this property. Compare to EDF, since the delay bound provided by GPS is not affected by the traffic characteristic or queue status of other flows, which makes the system more controllable and be able to provide QoS guarantee in per-flow basis.

b. Ideal flexibility

By varying the weight factors, we can enjoy the flexibility of treating the flows in a variety of different ways and providing widely different performance guarantees.

2. A packet-by-packet scheduling algorithm which can provide excellent approximation to GPS has been proposed [1]. This scheduling algorithm is known as packet-by-packet GPS (PGPS) or weighted fair queueing (WFQ). In the later section, we will discuss the operation and several important properties of PGPS in more detail.

3.6 Packet-by-packet Generalized Processor Sharing (PGPS)

PGPS is a scheduling algorithm which can provide excellent approximation to the ideal properties of GPS and is practical enough to be realized in a packet-switched system. The concept of PGPS is first proposed in [4] under the name Weighted Fair Queueing (WFQ). However, a great generalization and insightful analysis was done by Parekh and Gallager in the remarkable paper [1] and [2]. The basic idea of PGPS is simulating the transmission order of GPS system. More specific, let F_p be the time at which packet p will depart (finish service) under GPS system, then the basic idea of PGPS is to approximate GPS by serving

the packets in increasing order of F_p . However, sometimes there is no way for a work-conserving system to serve all the arrival packets in the exactly the same order as that of corresponding GPS system. To explain it, we make the following observations:

1. The busy period (the time duration that a server continuously sends packets) of GPS and PSPS is identical, since GPS and PGPS are all work-conserving system, the server will never idle and send packets with rate C when there are unfinished packets queued in the system.
2. When the PGPS server is available for sending the next packet at time τ , the next packet to depart under GPS may not have arrived at time τ . It's essentially due to the fact that a packet may depart earlier than the packets which arrive earlier than it under GPS. *A packet may arrive too late to be send in PGPS system*, at this time, if the system is work-conserving, the server should pick another backlogged packet to send, and this would conflict the sending order under GPS system. Since we do not have additional assumption to the arrival pattern of packets here, there is no way for the server to be both work-conserving and to always serve the packet in increasing order of F_p .

To preserve the property of work-conserving, the PGPS server picks the first packet that would complete service in the GPS simulation. In other words, if PGPS schedules a packet p at time τ before another packet p' that is also backlogged at time τ , then packet p cannot leave later than packet p' in the simulated GPS system.

We have known the basic operation of PGPS, now a natural question arises: how well does PGPS approximate GPS? To answer this question, we may attempt to find the worst-case performance under PGPS compared to that of GPS. So we ask another question: how much later packets may depart the system under PGPS relative to GPS? In fact, it can be proved that let the G_p be the time at which packet p departs under PGPS, then

$$G_p - F_p \leq \frac{L_{\max}}{C}$$

where L_{\max} is the maximum packet length. That is, the depart time of a packet under PGPS system is not later than that under GPS system by more than the time of transmitting one packet. To verify this result, we first present a useful property:

Lemma 1 *Let p and p' be packets in a GPS system at time τ and suppose that packet p complete service before packet p' if there are no arrivals after time τ . Then packet p will also complete service before packet p' for any pattern of arrivals after time τ*

Proof.

The flows to which packet p and p' belong are backlogged at time τ . By (3.2), the ratio of the service received by these flows is independent of future arrivals. ■

Now we have prepared to prove the worst-case delay of PGPS system.

Theorem 1 *For all packet p , let G_p and F_p be the departure time of packet p under PGPS and GPS systems, respectively. Then*

$$G_p - F_p \leq \frac{L_{\max}}{C}$$

where L_{\max} is the maximum packet length, and C is the outgoing link capacity.

Proof.

As observed above, the busy periods of GPS and PGPS coincide, that is, the GPS server is in a busy period if and only if the PGPS server is in a busy period. Hence it suffices to prove

the theorem by considering one busy period. Let p_k be the k -th packet in the busy period to depart under PGPS and let its length be L_k . Also let t_k be the time that p_k depart under PGPS and u_k be the time that p_k departs under GPS. Finally, let a_k be the time that p_k arrives. It should be first noted that, if the sending order in a busy period under PGPS is the same as that under GPS, then it can be easily verified the departure time of the packets under PGPS system are earlier or equal to those under GPS system. However, since the busy periods of GPS and PGPS systems coincide, there are only two possible cases:

1. The departure times of all the packets under PGPS system in a busy period are all the same as those of corresponding GPS system.
2. If the departure times of some packets under PGPS system in a busy period are earlier than that of GPS, then there are also some packets with which the departure time are later than those of corresponding GPS system.

The second case implies that if there is a packet with which the departure time under PGPS system is later than the departure time of the corresponding GPS system, then the sending orders are not the same in the two systems in the busy period. According to the operation of PGPS, the difference of sending orders is only caused by some packets arrive too late to be transmitted in their order in GPS system. Thus, after these packets arrive, they may wait for the packets which should be sent later than them in GPS system to be served. Then, the additional delay caused.

Now we are clear that the only packets that have later departure time under PGPS system than under GPS system are those that arrive too late to be send in the order of corresponding GPS system. Based on this fact, we now show that:

$$t_k \leq u_k + \frac{L_{\max}}{C}$$

For $k = 1, 2, \dots$. Let p_m be the packet with the largest index that has earlier departure time than p_k under PGPS system but has later depart time under GPS system. That is, m satisfies

$$\begin{aligned} 0 < m &\leq k - 1 \\ u_m > u_k &\geq u_i \quad \text{for } m < i < k \end{aligned}$$

So packet p_m is send before packets p_{m+1}, \dots, p_k under PGPS, but after all these packets under GPS. If no such m exists then set $m = 0$. For the case $m = 0$, it direct lead to case 1 above, and $u_k \geq t_k$. For the case $m > 0$, packet p_m begins transmission at $t_m - L_m/C$, so from Lemma 1:

$$\min\{a_{m+1}, \dots, a_k\} > t_m - \frac{L_m}{C}$$

That is, p_{m+1}, \dots, p_{k-1} arrive and are served under GPS system after $t_m - L_m/C$. Thus

$$u_k \geq \frac{1}{C}(L_k + L_{k+1} + \dots + L_{m+1}) + t_m - \frac{L_m}{C}$$

Moreover, since

$$\frac{1}{C}(L_k + L_{k-1} + \dots + L_{m+1}) + t_m = t_k$$

we obtain the inequality

$$u_k \geq t_k - \frac{L_m}{C} \geq t_k - \frac{L_{\max}}{C}$$

which directly lead to the desired result. ■

It is worth to note that the guarantee of delay in PGPS system in Theorem 1 leads to the guarantee of per-flow throughput.

Theorem 2 For all times τ and flows i

$$S_i(0, \tau) - S'_i(0, \tau) \leq L_{\max}$$

where $S_i(a, b)$ and $S'_i(a, b)$ are the amount of flow i traffic served in the interval $[a, b]$, respectively.

Prove.

$$S_i(0, \tau) - L_{\max} \leq S_i(0, \tau - \frac{L_{\max}}{C}) \stackrel{(a)}{\leq} S'_i(0, \tau)$$

relation (a) comes from the fact that all the flow i packets transmitted before $\tau - L_{\max}/C$ under GPS system will always be transmitted before τ under PGPS system, which is the direct consequence of Theorem 1. ■

Let $Q_i(\tau)$ and $Q'_i(\tau)$ be the flow i backlog at time τ under GPS and PGPS system, respectively. Then it immediately follows from Theorem 2 that

Corollary 2.1 For all time τ and flow i

$$Q'_i(\tau) - Q_i(\tau) \leq L_{\max}$$

From the above results, we can see that PGPS provides quiet close approximation to GPS with the service curve never falls behind more than one packet length. This allows us to relate results for GPS to the packet-switched system in a precise manner. For more extensive analysis of PGPS, readers can refer to [1], [2], and [3].

4. Wireless packet scheduling algorithms

Recently, as various wireless technologies and systems are rapidly developed, the design of packet scheduling algorithms in such wireless environments for efficient packet transmissions has been a crucial research direction. Till now, a lot of wireless packet scheduling algorithms have been studied in many research papers. In the section, we will select four much more representative ones for illustrations in detail.

4.1 Idealized Wireless Fair Queueing (IWFQ) algorithm

The Idealized Wireless Fair Queueing (IWFQ) algorithm, proposed by Lu, Bharghavan, and Srikant [14] is one of the earliest representative packet scheduling algorithms for wireless access networks and to handle the characteristic of location-dependent burst error in wireless links. IWFQ takes an error-free WFQ service system as its reference system, where a channel predictor is included in the system to monitor the wireless link statuses of each flow

and determines the links are in either “good” or “bad” states. The difference between IWFQ and WFQ is that when a picked packet is predicted in a bad link state, it will not be transmit and the packet with the next smallest virtual finish time will be picked. The process will repeat until the scheduler finds a packet with a good state.

A flow is said to be *lagging*, *leading*, or *in sync* when the queue size is smaller than, larger than, or equal to the queue size in the reference system. When a *lagging* flow recovered from a bad link state, it must have packets with smaller virtual finish times, compare to other error-free flows’ packets. Thus, it will have precedence to be picked to transmit. So the compensation is guaranteed [15]. Additionally, to avoid unbounded amount of compensation starve other flows in good link state, the total lag that will be compensated among all *lagging* flows is bounded by B bits. Similarly, a flow i cannot lead more than li bits.

However, IWFQ does not consider the delay/jitter requirements in real-time applications. It makes no difference for different kind of applications, but in fact, non-real-time and delay-sensitive real-time applications have fundamental difference in QoS requirement, so always treat them identically may not be a reasonable solution. In addition, the choice of the parameter B reflects a conflict between the worst-case delay and throughput properties. Hence, the guarantees for throughput and delay are tightly coupled. In many scenarios, especially for real-time applications, decoupling of delay from bandwidth might be a more attractive approach [16]. Moreover, since the absolute priority is given to packets with the smallest virtual finish time, so a lagging flow may be compensated in a rate independent of its allocated service rate, violating the semantics that a larger guaranteed rate implies better QoS, which may be not desirable.

4.2 Channel-condition Independent packet Fair Queueing (CIF-Q) algorithm

The Channel-condition Independent packet Fair Queueing (CIF-Q) algorithm [17], proposed by Ng, Stoica, and Zhang. CIF-Q also uses an error free fair queueing algorithm as a reference system. In [17], Start-time Fair Queueing (SFQ) is chosen to be the core of CIF-Q. Similar to IWFQ, a flow is also classified to be *lagging*, *leading*, or *satisfied* according to the difference of the amount of service it have received to that of the corresponding reference system. The major difference between CIF-Q and IWFQ is that in CIF-Q the leading flows are allowed to continue to receive service at an average rate ar_i , where r_i is the service rate allocated to flow i and a is a configurable parameter. And instead of always choosing the packet with smallest virtual service tag like IWFQ, the compensation in CIF-Q is distributed among the *lagging* flows in proportion to their allocated service rates.

Compared with IWFQ, CIF-Q has better scheduling fairness and also has good properties of guaranteeing delay and throughput for error-free flows like IWFQ. However, the requirement of decoupling of delay from bandwidth is still not achieved by CIF-Q.

4.3 Improved Channel State Dependent Packet Scheduling (I-CSDPS) algorithm

A wireless scheduling algorithm employing a modified version of Deficit Round Robin (DRR) scheduler is called Improved Channel State Dependent Packet Scheduling (I-CSDPS), which is proposed by J. Gomez, A. T. Campbell, and H. Morikawa [18].

In DRR, each flow has its own queue, and the queues are served in a round robin fashion. Each queue maintains two parameters: Deficit Counter (DC) and Quantum Size (QS). DC can be regarded as the total credit (in bits or bytes) that a flow has to transmit packets. And

QS determines how much credit is given to a flow in each round. For each flow at the beginning of each round, a credit of size QS is added to DC . When the scheduler serves a queue, it transmits the first N packets in the queue, where N is the largest integer such that $\sum_{i=1}^N l_i \leq DC$, where l_i is the size of the i th packet in the queue. After transmission DC is decreased by $\sum_{i=1}^N l_i$. If the scheduler serves a queue and finds that there are no packets in queue, its DC is reset to zero.

To allow flows to receive compensation for their lost service due to link errors, I-CSDPS adds a compensation counter (CC) to each flow. CC to keep track of the amount of lost service for each flow. If the scheduler defers transmission of a packet because of link errors, the corresponding DC is decreased by the QS of the flow and the CC is increased by the QS . At the beginning of each round, $\alpha \cdot CC$ amount of credit is added to DC , and CC is decreased by the same amount, where $0 < \alpha \leq 1$.

Also, to avoid problems caused by unbounded compensation, the credit accumulated in a DC cannot exceed a certain value DC_{\max} . Similar to the parameter B in IWFQ, the choice of DC_{\max} also lead to the tradeoff between delay bound and the compensation for a flow lost its service. However, this bound is very loose and is in proportion to on the number of all active flows.

4.4 Proportional Fair (PF) algorithm

In the recent years, the two most well-known packet scheduling schemes for future wireless cellular networks are the maximum carrier-to-interference ratio (Max CIR) [26] and the proportional fair (PF) [27] schemes. Max CIR tends to maximize the system's capacity by serving the connections with the best channel quality condition at the expense of fairness since those connections with bad channel quality conditions may not get served. PF tries to increase the degree of fairness among connections by selecting those with the largest relative channel quality where the relative channel quality is the ratio between the connection's current supportable data rate (which depends on its channel quality conditions) and its average throughput. However, a recent study shows that the PF scheme gives more priority to connections with high variance in their channel conditions [28]. Therefore, we pay our attention focusing on the PF scheme for illustration here.

In another point of view, in wireless communication systems, the optimal design of forward link gets more attention because of the asymmetric nature of multimedia traffic, such as video streaming, e-mail, http and Web surfing. For the efficient utilization of scarce radio resources under massive downlink traffic, opportunistic scheduling in wireless networks has recently been considered important.

The PF was originally proposed in the network scheduling context by Kelly *et al.* in [45] as an alternative for a max-min scheduler, a PF scheduling promises an attractive trade-off between the maximum average throughput and user fairness.

The standard PF scheme in packet scheduling was formally defined in [45].

Definition: A scheduling P is 'proportional fair' if and only if, for any feasible scheduling S , it satisfies:

$$\sum_{i \in U} \frac{R_i^{(S)} - R_i^{(P)}}{R_i^{(P)}} \leq 0$$

where U is the user set and $R_i^{(S)}$ is the average rate of user i by scheduler S .

Also, it is known that a PF allocation P should maximize the sum of logarithmic average user rates [21], which is expressed by

$$P = \arg \max_S \sum_{i \in U} \log R_i^{(S)} .$$

The PF scheduling is implemented for Qualcomm's HDR system, where the number of transmission channels is one. Only one user is allocated to transmit at a time, and the PF is achieved by scheduling a user j according to

$$j = \arg \max_i \frac{r_i}{\bar{R}_i} ,$$

where r_i is the instantaneous transmittable data rate at the current slot of user i and \bar{R}_i is the average data rate at the previous slot of user i .

Consider a model where there are N active users sharing a wireless channel with the channel condition seen by each user varying independently. Better channel conditions translate into higher data rate and vice versa. Each user continuously sends its measured channel condition back to the centralized PF scheduler which resides at the base station. If the channel measurement feedback delay is relatively small compared to the channel rate variation, the scheduler has a good enough estimate of all the users' channel condition when it schedules a packet to be transmitted to the user. Since channel condition varies independently among different users, PF exploits user diversity by selecting the user with the best condition to transmit during different time slots.

The PF algorithm was proposed after studying the unfairness exhibited when increasing the capacity of CDMA by means of differentiating between different users. Transmission of pilot symbols to the different users yields channel state information, and by allocating most resources to the users having the best channels, the total system capacity of the CDMA scheme could be increased. Such allocation of resources favors the users closest to the transmitting node, resulting in reduced fairness between the different users. The PF algorithm seeks to increase the fairness among the users at the same time as keeping some of the high system throughput characteristics.

The PF scheduling algorithm has received much attention due to its favorable trade-off between total system throughput and fairness in throughput between scheduled users [19] [20]. The PF scheduling algorithm can achieve multi-user diversity [20] [21], where the scheduler tracks the channel fluctuations of the users and only schedules users when their instantaneous channel quality is near the peak. In other words, the PF scheme is a channel-state based scheduling algorithm that relies on the concept of exploiting user diversity.

PF has extensively been studied under well-defined propagation channel conditions, such as flat fading channels with Rayleigh and/or Rician type of fading [22], or the ITU Vehicular and Pedestrian channels [24], which are typically applied in standardization work [23].

In early years, the PF scheduling is widely considered in single-carrier situations. In addition, it is pointed out in [26] that the PF scheme for a single antenna system is attractive for non-real time traffics, since it achieves substantially larger system throughput than the Round-Robin (RR) scheme. The PF scheme also provides the same level of fairness as the RR

scheme in the average sense [25]. Further descriptions of the PF algorithm can be in [29], [30], [31], [32] and [33], while a variant which offers delay constraints is described in [34]. In more recent years, as many modern broadband wireless systems with multi-carrier transmissions are rapidly developed, multi-carrier scheduling becomes a hot topic. The issue will be investigated and illustrated in detail in Section 4.

5. Case study: design of packet scheduling schemes for OFDMA-based systems

5.1 Introduction to OFDMA

Recently there has been a high demand for large volume of multimedia and other application services. Such a demand in wireless communication networks requires high transmission data rates. However, such high transmission data rates would result in frequency selective fading and Inter-Symbol Interference (ISI). As a solution to overcome these issues, Orthogonal Frequency Division Multiplexing (OFDM) had been proposed in [35].

Nowadays, the OFDM technology has widely been used in most of the multi-user wireless systems, which can be referred to research papers [36-38] for instance. When such a multiple carrier system has multi-user, it can be referred to as Orthogonal Frequency Division Multiple Access (OFDMA) system. In other words, the key difference between both transmission methods is that OFDM allows only one user on the channel at any given time whereas OFDMA allows multiple accesses on the same channel. OFDMA assigns a subset of subcarriers to individual users and their transmissions are simultaneous. OFDMA functions essentially as OFDM-FDMA. Each OFDMA user transmits symbols using some subcarriers that remain orthogonal to those of other users. More than one subcarrier can be assigned to one user to support high data rate applications. Simultaneous transmissions from several users can achieve better spectral efficiency.

5.2 Token-based packet scheduling scheme for IEEE 802.16 [46]

5.2.1 Frame by frame operation scheme

Since IEEE 802.16 is a discrete-time system, time is divided into fixed-length frames, and every MS is mandatory to synchronize with the BS before entering the IEEE 802.16 network [45], our packet scheduler scheme is also a discrete-time scheme and schedule packets in a per-frame basis. Additionally, because we consider downlink traffic only, all the components and algorithms are all operated in BS.

Figure 5.1 is a simple description of the operation of our packet scheduling scheme. When a packet arrives at the BS from the upper layer, it is buffered in the BS first and the system decides whether it will be scheduled to be transmitted in the next frame. This procedure will be repeated every frame until this packet is transmitted successfully in the downlink subframe of one of the afterward frame. We assumed that a packet transmitted in the downlink subframe of a frame will receive ARQ feedback (ACK or NAK) immediately from MSs in the uplink subframe of the same frame. The result of scheduling of the next frame is broadcast via the DL-MAP which is transmitted at the beginning of the next frame.

1. System Resource Normalization

Since the packets of each flow may be transmitted in different Modulation and Coding Schemes (MCS), we use "slots" as a general unit of entire system to describe traffic characteristic and system resource. Suppose that the MCS used for a flow is not changed during the session's life time.

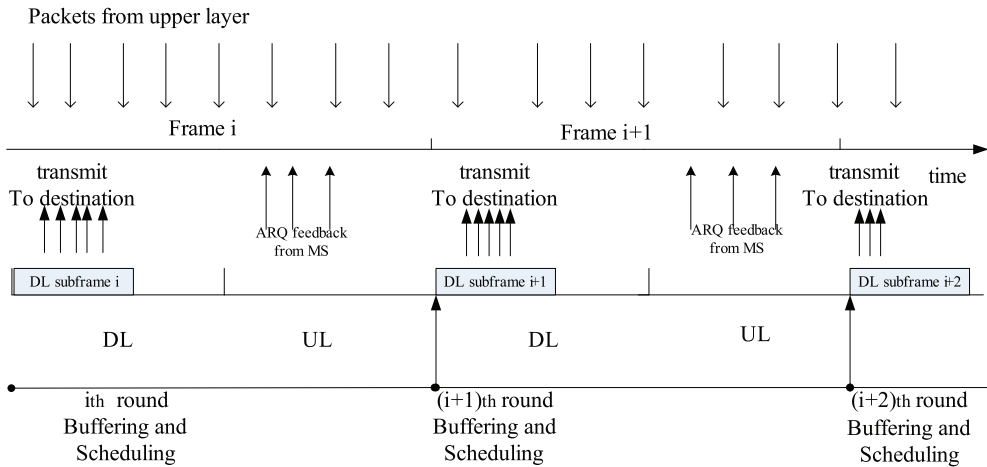


Fig. 5.1 Simple description of the operation of packet scheduling scheme

For example, we can say a leaky bucket shaper has bucket depth 10 slots and 2.251 slots/frame. Or we guaranteed a non-real-time session a minimum throughput of 5.35 slots/frame.

In this study, we assume our system has a total of C slots available for downlink traffic in a frame. We can also say that this system has a capacity of C slots/frame.

2. System Architectures

Figure 5.2 is our proposed packet scheduling scheme operated in the BS. It consists of several components. We describe their functions and algorithms respectively in this section. *Classifier and Traffic Profile*

The classifier is responsible for classifying packets from upper layers to the appropriate service group. Two service groups are defined, that is, real-time group and non-real-time group, according to their fundamental differences of QoS requirement. There are several approaches to identify each packet's group. One suggestion is to classify each packet according to the service type of its MAC connection ID. For example, UGS, rtPS, and ertPS are belong to real-time group and nrtPS and BE are belong to non-real-time group. We also assume that each flow has a flow profile for description of its traffic characteristic. Flows of real-time group and flows of non-real-time group have different flow profiles. We introduce them respectively as follows:

Real-time group: Real-time flows are delay-sensitive traffic. A packet from real-time sessions is expected to be transmitted successfully in some delay constraint or it is regarded as meaningless and dropped. Although that, some loss rate does not degrade the application layer quality seriously and is tolerable for users. A triple $\{\delta_i, \lambda_i, D_i\}$ is used to describe the traffic characteristic of real-time flow i . Where δ_i is maximum burst size (normalized to slots), λ_i is the minimum sustainable data rate (normalized to slots/frame), D_i is the maximum tolerable packet delay (in frame). Note that when a leaky bucket policer [46] is used, δ_i is equivalent to the bucket depth and λ_i is equivalent to the average rate in the leaky-bucket policing algorithm.

Non-real-time group: Non-real-time flows are not sensitive to delay and jitter. The QoS matrix of non-real-time services is the average throughput. A parameter λ_j is used to describe the traffic characteristic of non-real-time session j . Where λ_j is the minimum reserved data rate (normalized to slots/frame).

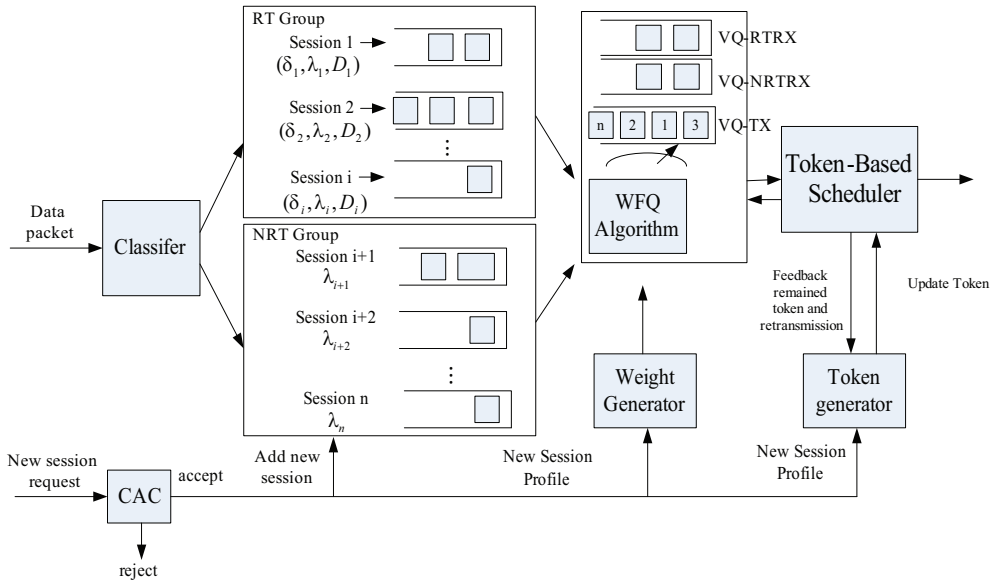


Fig. 5.2 The system architecture of our packet scheduler

Packet Scheduler and Weight Generator

In the following sections, we introduce the main part of our packet scheduling scheme. Some useful notations are as follows:

S_{RT} : The set of all real-time flows

S_{NRT} : The set of all non-real-time flows

b_i : The minimum required capacity to achieve the QoS requirement of flow i (normalized to slots/frame). For real-time services, the QoS matrix is the tolerant delay of a packet, and for non-real-time services, the QoS matrix is the average throughput

w_i : The *weighting factor* of flow i in the WFQ scheduler

t_i : The current token value of flow i . If a packet of flow i is scheduled to transmit in the next frame. It must take the token value equal to the size of the packet (normalize to slot) away

T_i : The maximum token value flow i can keep

α_i : The protecting factor of flow i . A number which is larger than or equal to 1. The more α_i is, the more *protected capacity* for flow i .

r_i : The token incremental rate of flow i . At the beginning of a frame, the token value of flow i is updated to $\max(t_i + r_i, T_i)$. r_i can be regard as the *protected capacity* for flow i . $r_i = b_i * \alpha_i$.

R : The sum of the *protected capacity* of real-time flows, $R = \sum_{i \in S_{RT}} r_i$

N : The sum of the *protected capacity* of non-real-time flows, $N = \sum_{i \in S_{NRT}} r_i$

N_{max} : The maximum value of the sum of *protected capacity* of non-real-time flows

C : The available slots for downlink traffic per frame. Intuitively, $C \geq R + N$

Our packet scheduling scheme has two stages. The first stage is a work conserving packet scheduler. When a packet arrives from upper layer, it first enters the first stage. The actual conditions in the lower layer such as the channel status or the allocation of slot are transparent to the first stage. It always assumes there is an error-free channel with fixed capacity C in the lower layer. The main purpose of the first stage packet scheduler is to emulate the transmission order of a work conserving system in an ideal condition and be a reference system to our scheme. The packet order in the reference system is not certainly the actual transmission order in our packet scheduling scheme. The task of determining which packet should be scheduled to transmit in the next frame is executed by the token-based slot scheduler which is in the second stage of our scheme. The detail of the operation of the token-based slot scheduler will be illustrated in the next section.

There is no constraint to the scheduling discipline adopted in the first stage packet scheduler. But to achieve a better resource allocation and isolation among each flow, weight-based scheduling disciplines such as WFQ, VC, are suggested. In our packet scheduling scheme, we take WFQ as the reference system. When a packet arrives from the upper layer, it enters the packet scheduler in the first stage, the packet scheduler then schedules the transmission order of this packet with WFQ algorithm. The packet order scheduled by the packet scheduler is recorded in a *virtual queue*. *Virtual queue* is not really buffered the packets but store the pointers of packet which is the input of the token-based slot scheduler in the second stage.

There are three virtual queues with strict priorities. They are virtual queue for real-time retransmission (VQ-RTRX), virtual queue (VQ-NRTRX) for non-real-time retransmission, and virtual queue for first time transmission (VQ-TX) according to their priorities. The packets which have not been transmitted are recorded their pointer in the VQ. The real-time packets which have transmitted but not received successfully by the receiver, their pointers are moved from VQ to VQ-RTRX. The non-real-time packets which have transmitted but not received successfully by the receiver, their pointers are moved from VQ to NRTRVQ. The token-based packet scheduler checks the packet pointers from the virtual queue with highest priority (RTRVQ) to that with lowest priority (VQ) and determines which packets will be scheduled to transmit in the next frame. The algorithm determining which packets will be scheduled will be discussed in the next section in detail. Figure 5.3 is the queueing model of our packet scheduling scheme.

To indicate the resource sharing of the flows, each flow i associates a *weighting factor* w_i . The *weighting factor* is an important parameter as the weight in packet scheduler in the first stage and in the *debt allocation procedure* in token-based slot scheduler. We will return to discuss the procedure of weight allocation after we introduce the token-based slot scheduler and its algorithm in the next section.

Token-Based Scheduler

We use a token-based scheduler to determine which packet should be scheduled to transmit in the next frame. The fundamental operation of the token-based slot scheduler is as follows. For convenience of illustration, we define some notation as follows:

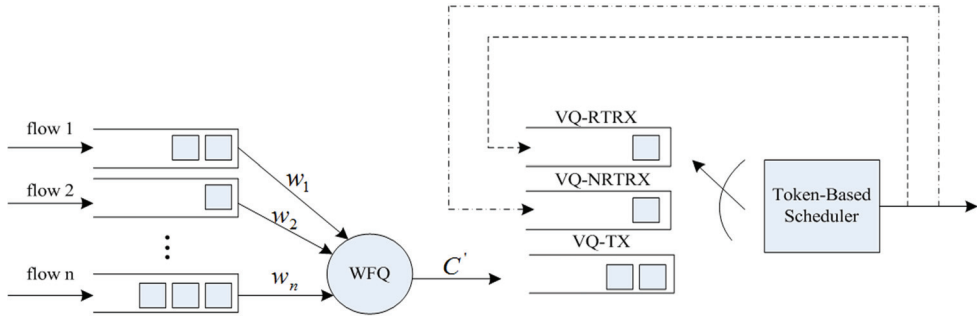


Fig. 5.3 The queuing model of our proposed packet scheduling scheme

l : The size of the packet that is checked by the token-based scheduler

l_i^j : The j th packets of flow i which is scheduled in the next frame

L_i : The total length of the scheduled packets of flow i . That is, $L_i = \sum_j l_i^j$

$remained_slots$: The remained slots available for scheduling. $remained_slot = C - \sum_i \lceil L_i \rceil$

Assume that there are C slots available for downlink traffic each frame. Each flow i maintains a *token value* t_i . The *token value* of every session i has a fixed *token incremental rate* r_i , the unit of r_i is slots/frame. At every beginning of a frame, the *token value* of session i is increased by r_i slots. The task of updating the token value of each flow at the beginning of a frame is operated by the *token generator*. When a packet of flow i with size l (normalized to slots) is scheduled to transmit, it must take the token value equal to the amount of the size of the packet (normalized to slots) away. And the number of slots available for scheduling is also decreased by That is,

When a packet of flow i with size l is scheduled

$$t_i \leftarrow t_i - l \tag{5.1}$$

$$L_i += l \tag{5.2}$$

$$remained_slot = C - \sum_i \lceil L_i \rceil \tag{5.3}$$

There should be an upper bound of the token value t_i , where we denote it by T_i . The setting of T_i can affect the system performance. We will discuss the issue of the effect of T_i later.

Thus, when a new frame start

$$t_i \leftarrow \max(t_i + r_i, T_i), \text{ for each flow } i \tag{5.4}$$

We can regarded r_i as the *protected capacity* of flow i . The configuration of r_i can affect the system performance significantly. We introduce the detailed algorithm of token-based slot scheduler in this section. Then we will return to discuss the guideline of the setting of token incremental rate in the next section.

The basic principle of scheduling is as follows:

1. The packet which has sufficient token value to transmit it has the higher priority, or it has the lower priority
2. For the packets with the same priority, the scheduling order is according to the order in WFQ scheduler (i.e. the order of the virtual finish time in the WFQ)

The process of our token-based packet scheduler algorithm can be divided into two phases. At the beginning of scheduling, the token-based packet scheduler enters the first phase. It checks the packet pointers sequentially in each virtual queue from high priority to low priority. We call the first step *packet selection procedure*. If the checked packet has sufficient token value (that is, $t_i \geq l$) and there are sufficient slots to transmit it in the next frame (that is, $remained_slots \geq \lceil L_i + l \rceil - \lfloor L_i \rfloor$). It is scheduled in the next frame. And the token value of this flow is decreased by the size of the packet.

Otherwise, the token-based slot scheduler will skip it, and to keep the packets of the same flow to be transmitted in order, other packets from the same flow which have not been checked are also skipped in the first phase. For convenience of discussion, we say that this flow is *blocked* in this phase. After all the packets are checked, the slot scheduler enters the second phase.

During the second phase, the slot scheduler continues to find other packets can be transmitted with the remained slots in the next frame. The token-based scheduler does it by checking the packets which have not been scheduled in the first phase. Again, the order of checking is the same as the first phase. If there are still sufficient slots to transmit the checked packet (that is, $remained_slots \geq \lceil L_i + l \rceil - \lfloor L_i \rfloor$), the packet is scheduled in the next frame, or the packet is skipped and the flow of this packet is *blocked* which is the same as the first phase. When the checked packet is scheduled, the token value of the scheduled packet must run out and become a negative number, it implies that the session of this packet uses more capacity than its protected capacity. The additional consumed token value exceeding the *protected capacity* is regarded as the *debt* draw from other flows. For example, if $t_i=30$, now flow i has a packet with size 50 slots and is scheduled to transmit in the next frame by the scheduler. The *debt* is 20. If $t_i=-10$, a packet with the same size is scheduled, the *debt* is 50.

We can represent *debt* as follows:

$$debt = -1 * \max(-l, t_i - l), \quad l \geq t_i \quad (5.5)$$

In our algorithm, we prefer to give real-time sessions more opportunity to increase its token value, since if we clean the packets of real-time sessions as soon as possible, it is more likely to have more remained resource to improve the throughput of non-real-time traffic in the operation of our token-based algorithm, thus meet the QoS requirement of both. So the *debt* is allocated to the token value of all real-time flows in proportion to their weight. That is,

$$t_i = \max(T_i, t_i + \frac{w_i}{\sum_{j \in S_{RT}} w_j} * debt), \quad \text{for all } i \in S_{RT} \quad (5.6)$$

We call the second step *debt allocation procedure*. For example, there are three flows. Flow 1 and 2 is real-time flows with *weighting factor* 0.3 and 0.2 respectively, flow 3 is non-real-time flows with *weighting factor* 0.5. And their *token value* is -10, 40, 20. Now flow 2 has a packet of size 50 slots be scheduled by the token-based slot scheduler. Since the packet size is larger

than the token value of flow 2. We can calculate the *debt* is 10 and allocate it to the token value of real-time flows, that is, flow 1 and flow 2. Finally, the token value of flow 1 is $-10 + \frac{10 * 0.3}{0.2 + 0.3} = -4$, the token value of flow 2 is $40 - 50 + \frac{10 * 0.2}{0.2 + 0.3} = -6$. The token value of flow 3 is not changed. The debt allocation procedure is finished when all the packets are checked. Then in the slot the scheduler transmits the scheduled packet and receives ARQ from the receivers.

The chosen of T_i can affect system performance significantly. If the T_i is set too large, suppose flow i is in good channel status for a long time and it accumulates a large amount of token value from the token generator and the *debt* of other flows, now it incurs burst error and the channel is in bad channel for an long interval of time. Then flow i will waste a large amount of system resource to transmit error packet because it accumulates too much token value when it is in good channel. This is unfavorable. On the other hand, if the maximum token value is set too small. Then it is hard to differentiate the flows behave well and give it more opportunity to be scheduled. Thus is difficult to show the advantage of our algorithm. Furthermore, the traffic characteristic also should be taken into consider. Generally, we suggest that the maximum token value of non-real-time flows has better larger than that of real-time flows, because most non-real-time flow are TCP traffic, which is composed of several burst.

The flow chart of all procedures of the token-based slot scheduler is shown in Fig. 5.4. The checking procedure and the debt allocation procedure is the core of our slot scheduler. The pseudo codes of these two procedures are shown in Fig. 5.5 and Fig. 5.6, respectively.

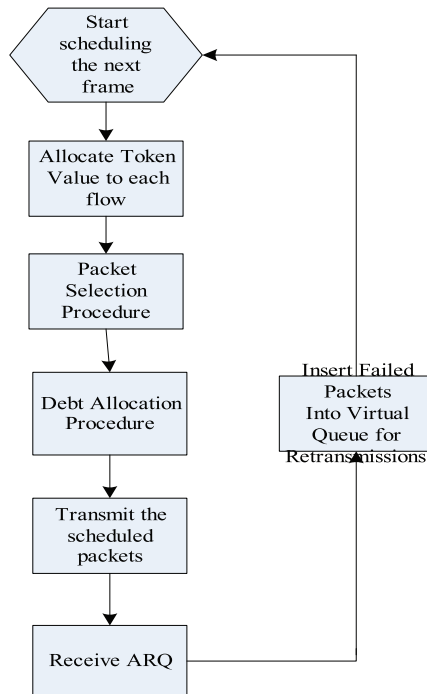


Fig. 5.4 The flow chart of the procedure of the token-based scheduler

```

packet-selection-procedure(system_capacity){
  remained_slot ← system_capacity;
  for(each flow j){
    blockj=0;
  }
  for(each virtual queue, from highest priority to lowest priority)
  while(packets not checked in the virtual queue){
    i ← the flow the packet belong to;
    l ← the size of the checked packet;
    if( remained_slot ≥ ⌈Li + l⌉ - ⌊Li⌋ and l ≤ ti and blocki≠0){
      schedule this packet in the next frame;
      remained_slot = ⌈Li + l⌉ - ⌊Li⌋;
      ti = ti - l;
    }
    else
      blocki=1; /* the other packet of flow i is also skipped in this procedure */
  }
  }
  debt-allocation-procedure(remained_slot);
}

```

Fig. 5.5 The pseudo code of packet selection procedure

Weight Allocation and Token Generator

In this section, we discuss the functions of *token generator*, and the relationship between . The main tasks of token generator are calculating the *token incremental rate* of each flow, and allocating token value to all the flows per frame. In this section, we address the issues of the chosen of *weighting factor* and *token incremental rate*.

The different configuration of the *token incremental rate* alters the system performance. We suggest that the *token incremental rate* of flow i is set to its *minimum required capacity* b_i multiplies a *protecting factor* α_i . The *minimum required capacity* of flow i is the minimum capacity need to reserved for flow i to satisfy its QoS requirement. That is, the capacity to make flow i 's QoS acceptable in the assumption that no channel error occurs. For real-time services, the QoS matrix is the tolerable delay of a packet. A real-time flow i with traffic profile $\{\delta_i, \lambda_i, D_i\}$, the *minimum required capacity* is $\max(\frac{\delta_i}{D_i}, \lambda_i)$. For non-real-time services, the QoS matrix is the average throughput. A non-real-time flow j with traffic profile λ_j , the *minimum required capacity* is λ_j . Thus, b_i is calculated as follows:

$$b_i = \begin{cases} \max(\frac{\delta_i}{D_i}, \lambda_i), & i \in S_{RT} \\ \lambda_i, & i \in S_{NRT} \end{cases} \quad (5.7)$$

The purpose of *protecting factor* α_i is to expand the *protected capacity* of flow i by multiplying the *minimum required capacity* by a number larger than or equal to 1. The larger the *protecting*

```

debt_allocation_procedure(remained_slot){
  for(each flow j)
    blocki ← 0;

  for(each virtual queue, from the highest priority to the lowest priority){

    while(packets not checked and not scheduled in the virtual queue){
      i ← the flow the packet belong to;
      l ← the size of the packet;
      if (remained_slot ≥ ⌈Li + l⌉ - ⌊Li⌋ and blocki ≠ 1){
        schedule the packet in the next frame;
        debt ← -1 * max(ti - l, -l);
        ti - = l;
        remained_slot - = ⌈Li + l⌉ - ⌊Li⌋;
        for(all real-time flows k){
          
$$t_k = \max\left(t_k + \frac{w_k * debt}{\sum_{j \in RT} w_j}, T_k\right);$$

        }
      }
      else
        blocki ← 1; /* the other packet of flow i is also skipped in this procedure */
    }
  }
}

```

Fig. 5.6 The pseudo code of debt allocation procedure

factor, the larger the *protected capacity*. It implies that providing a flow more protection by giving it more resource than it required to compensate the loss due to wireless channel error. The tuning of *protecting factors* is also important and closely relative to system performance. If the *protecting factor* of a flow is too large, it may be unfair to other flows and also cause waste of resource, which will be harmful to overall system performance. In our scheme, we set the *protecting factors* of real-time flows to 1, and set the *protecting factors* of non-real-time flows to a number slightly larger than 1, for example, Since in our token-based scheduler, we give real-time flows more opportunity to increase their token value than that of non-real-time flows by allocating all the *debt* to real-time flows. So we compensate non-real-time flows by regulating their *protecting factors* to be larger than that of real-time flows'. Additionally, setting the *protecting factor* of real-time flows to 1 means the *protected capacity* of a real-time flow is the same as its *minimum required capacity*. It brings benefits to differentiate the flows with good channel status and the flows with bad channel status. Because when a flow suffers burst error, it will use more capacity than its minimum required, so it soon runs out of its token value, and other real-time flows in good channel status get additional token value. It makes the real-time flows in good channel status has higher priority to be transmitted, and improve the efficiency of the use of system resource.

In many real situations, we may degrade some resource sharing of non-real-time flows to make the system to accommodate more real-time flows. That is, satisfy more real-time users at the cost of some average throughput of non-real-time services. We can achieve this by bounding the sum of the *token generating rate* of non-real-time flows. When the sum of the *token generating rate* of all non-real-time flows exceeds a defined value N_{\max} , the *token generating rate* of all non-real-time flows degrade proportionally to make their sum not larger than N_{\max} .

Thus, the sum of the token rate of all non-real-time flows N can be represented as

$$N = \min(N_{\max}, \sum_{i \in S_{NRT}} (\alpha_i * b_i)) \quad (5.8)$$

and the token rate of a non-real-time flow can be calculated as

$$r_i = N * \frac{b_i}{\sum_{i \in S_{NRT}} b_i} \quad (5.9)$$

The *weighting factor* is for indicating the resource allocation of the WFQ in the first stage of our scheme. The WFQ emulates a work-conserving system with error-free channel. The *weighting factor* of flow i is proportional to its *protected capacity* divided by its protecting factor. That is,

$$w_i = \frac{r_i / \alpha_i}{\sum_j r_j / \alpha_j} \quad \text{for all flow } i \quad (5.10)$$

5.3 PF schemes for OFDMA systems

Recently, for the higher rate data transmission, interest in wireless communications has shifted in the direction of broadband systems such as multicarrier transmission systems such like the OFDMA system. There has been a growing interest in defining radio resource allocation for a physical layer based on the OFDMA technology for 4G cellular system. While throughput-optimal scheduling can be achieved by using the multi-user diversity effect, it can generate unfairness as users with bad channel conditions have a lower probability to get a resource.

Based on the definition of the standard PF scheduling scheme [45], the theorem of the modified PF scheduling for multi-carrier transmission systems was proposed in [40]. Notice that the proof of this theorem is omitted and can be referred to [40].

Theorem: A scheduling P is 'proportional fair' for a multicarrier transmission system, if and only if, for any feasible scheduling S , it satisfies:

$$P = \arg \max_S \prod_{i \in U} \left(1 + \frac{\sum_{k \in C_i} r_{i,k}}{(T-1)\bar{R}_i} \right),$$

where U is the set of selected users by S , C_i is the set of carriers allocated to user i , $r_{i,k}$ is the instantaneous transmittable data rate of carrier $k \in C_i$ at the current slot, \bar{R}_i is the average rate of user i at the previous slot, and T is the average window size.

With OFDMA, there are multiple transmission channels that can be used, where scheduling schemes considering the PF algorithms have widely been studied in many papers. See, for example, [39 41-42]. Papers [39] and [42] had proposed heuristic approaches by simply applying PF of the single carrier case in each subcarrier to adapt for the multi-carrier case in a suboptimal manner. Additionally the QoS requirement for each user was considered in [41]. Furthermore, readers are suggested to refer to [43-44] for more complete investigation of related modified PF schemes in multi-carrier systems.

6. Summary and the discussion of open issues

Packet scheduling is one of most important radio resource management functions. It is responsible for determining which packet is to be transmitted such that the resources are fully utilized. The design of an efficient algorithm to be used for the scheduling of packet transmissions in wireless communication networks is still an open issue for research. This Chapter has widely covered the conceptual description of many representative packet scheduling algorithms deployed in high-speed point-to-point wireline and wireless scenarios. Well designing algorithms with low complexity offering fairness among and potentially differentiation between different data-flows is important in the evolution of communication networks. The rapidly growing demand of network nodes capable of taking into account the different QoS requirements of different flows to better utilize the available resources at the same time as some degree of fairness is maintained, makes more intelligent packet scheduling a central topic in future development of communication technologies.

7. Acknowledgement

The authors wish to express their sincere appreciation for financial support from the National Science Council of the Republic of China under Contract NSC 98-2221-E-002-002.

8. References

- [1] A. Parekh, R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," *IEEE/ACM Trans. on Networking*, Vol. 1, June 1993
- [2] A. Parekh, R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The multiple-node case," *IEEE/ACM Trans. On Networking*, Vol. 2, April 1994
- [3] R. Cruz, "Quality of Service Guarantees in Virtual Circuit Switched Networks," *IEEE J. Select. Areas Commun.*, Special issue on "Advances in the Fundamentals of Networking," August, 1995.
- [4] A. Demers, S. Keshav, and S. Shenkar, "Analysis and simulation of a fair queueing algorithm," *Internet. Res. And Exper.*, vol. 1, 1990
- [5] D. Ferarri, "Real-time communication in an internetwork," *J. High Speed Networks*, vol. 1, no. 1, pp. 79-103, 1992

- [6] D. Ferrari and D. Verma, "A scheme for real-time channel establishment in wide-area networks," *IEEE Journal on Selected Areas in Communications*, 8(3):368-379, April 1990.
- [7] J. R. Piney, S. Sallent, "Performance Evaluation of a Normalized EDF Service Discipline," in *Proc. IEEE MELECON 2004, Dubrovnik, Croatia, May 2004*
- [8] S. Chaudhry, and A. Choudhary, "Tune dependent priority scheduling for guaranteed QoS Systems," in *Proc. Sixth International Conference on Computer Communications and Networks*, pp. 236-241, Sept. 1997
- [9] L. Georgiadis, R. Guerin, A. Parekh, "Optimal multiplexing on a single link: delay and buffer requirements," *IEEE Trans. On Information Theory*, 43(5), pp. 1518-1535, Sep. 1997
- [10] K. Zai, Y. Zhang, Y. Viniotis, "Achieving end-to-end delay bounds by EDF scheduling without traffic shaping," in *Proc. Infocom'01*, 2001
- [11] V. Sivaraman, F. M. Chiussi, Mario Gerla, "End-to-End Statistical Delay Service under GPS and EDF Scheduling: A Comparison Study," in *Proc. Infocom'01*, 2001
- [12] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, vol. 83, No. 10, Oct. 1995
- [13] J. Liebeherr, D. Wrege and D. Ferrari, "Exact admission control in networks with bounded delay services," *IEEE/ACM Trans. Networking*, vol. 4, pp. 885-901, 1996.
- [14] S. Lu and V. Bharghavan, "Fair Scheduling in Wireless Packet Networks," *IEEE/ACM Trans. Networking*, vol. 7, no. 4, pp. 473-489, 1999.
- [15] Y. Cao, and VICOR O. K. Li, "Scheduling Algorithms in Broad-Band Wireless Networks," *Proceedings of the IEEE*, Vol. 89, No. 1, Jan. 2001.
- [16] F. Tsou, H. Chiou, and Z. Tsai, "WDFQ: An Efficient Traffic Scheduler with Fair Bandwidth Sharing for Wireless Multimedia Services," *IEICE TRANS. COMMUNICATIONS*, Vol. E00-A, No. 1, Jan. 2000
- [17] T. S. Eugene Ng, I. Stoica, and H. Zhang, "Packet Fair Queueing Algorithms for Wireless Networks with Location-Dependent Errors," in *Proc. INFOCOM'98*, Mar. 1998, pp. 1103-1111.
- [18] J. Gomez, A. T. Campbell, and H. Morikawa, "The Havana Framework for Supporting Application and Channel Dependent QOS in Wireless Networks," in *Proc. Seventh International Conference on Network Protocols*, 1999.
- [19] T. E. Kolding, K. I. Pedersen, J. Wigard, F. Frederiksen, and P. E. Mogensen, "High Speed Downlink Packet Access: WCDMA Evolution," *IEEE Vehicular Technology Society News*, February 2003, pp. 4-10.
- [20] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beam forming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, June 2002.
- [21] D. Tse, "Multiuser diversity in wireless networks," *Wireless Communication Seminar*, Stanford University, April 2001.
- [22] J. M. Holtzman, "Asymptotic analysis of Proportional Fair algorithm," *IEEE Proc. Personal Indoor Mobile Radio Communications (PIMRC)*, September 2001, pp. 33-37.
- [23] T. E. Kolding, "Link and system performance aspects of Proportional Fair scheduling in WCDMA/HSDPA," *Proceedings of 58th IEEE Vehicular Technology Conference (VTC)*, Florida USA, October 2003, pp. 1454-1458.
- [24] "Guidelines for the evaluation of radio transmission technologies for IMT-2000," Recommendation ITU-R M.1225, 1997.

- [25] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE Veh. Technol. Conf. Spring*, Tokyo, Japan, May 2000, pp. 1854-1858.
- [26] S. Borst, "User-level performance of channel-aware scheduling schemes in wireless data networks," *IEEE INFOCOM*, Mar. 2003, vol. 1, pp. 321-331.
- [27] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE VTC*, May 2000, pp. 1854-1858.
- [28] M. Kazmi and N. Wiberg, "Scheduling schemes for HSDSCH in a WCDMA mixed traffic scenario," in *Proc. IEEE Int. Symp. PIMRC*, Beijing, China, Sep. 2003, pp. 1485-1489.
- [29] J. M. Holtzman, "Asymptotic Analysis of Proportional Fair Algorithm", *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Vol. 2, October 2001
- [30] L. Erwu, and K. K. Leung, "MAC 20-5 - Proportional Fair Scheduling: Analytical Insight under Rayleigh Fading Environment", *IEEE Wireless Communications and Networking Conference*, April 2008
- [31] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic Beamforming using Dumb Antennas", *IEEE Transactions on Information Theory*, Vol. 48, Issue 6, June 2002
- [32] P. Mueng, W. Yichuan, and W. Wenbo, "Joint an Advanced Proportionally Fair Scheduling and Rate Adaptation for Multi-services in TDD-CDMA Systems", *IEEE 59th Vehicular Technology Conference*, Vol. 3, May 2004
- [33] K. Kuenyoung, K. Hoon, and H. Youngnam, "A Proportionally Fair Scheduling Algorithm with QoS and Priority in 1xEV-DO", *IEEE Symposium on Personal, Indoor and Mobile Radio Communications*, Vol. 5, September 2002
- [34] O. S. Shin, and K. B. Lee, "Packet Scheduling over a Shared Wireless Link for Heterogeneous Classes of Traffic", *IEEE International Conference on Communications*, Vol. 1, June 2004
- [35] J. A.C. Bingham, "Multi carrier modulation for data transmission: an idea whose time has come," *IEEE Communications Magazine*, pp. 5-14, May 1990.
- [36] J. Jang and K. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE Journal on Selected Areas in Communications*, 21(2): 171-178, Feb. 2003.
- [37] Y. J. Zhang and K. B. Letaief, "Multiuser adaptive subcarrier-and bit allocation with adaptive cell selection for OFDM systems," *IEEE Transactions on Wireless Communications*, 3(4): 1566-1575, Sept. 2004.
- [38] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation for multiuser OFDM with constrained fairness," *IEEE Transactions on Wireless Communications*, 4(6): 2726-2737, Nov. 2005.
- [39] W. Anchun, X. Liang, Z. Shidong, X. Xibin, and Y. Yan, "Dynamic resource management in the fourth generation wireless systems," in *Proc. ICCT*, vol. 2, April 2003, pp. 1095-1098.
- [40] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Communications Letters*, vol. 9, no. 3, pp. 210-212, March 2005.
- [41] Y. Lu, C. Wang, C. Yin ,and G. Yue, "Downlink scheduling and radio resource allocation in adaptive OFDMA wireless communication system for user-individual QoS," *International Journal of Electrical, Computer, and Systems Engineering* 2009

- [42] N. Ruangchaijatupon and Y. Ji, "Simple proportional fairness scheduling for OFDMA frame-based wireless system," *IEEE WCNC 2008*
- [43] M. Kaneko, P. Popovski, and J. Dahl, "Proportional fairness in multi-carrier system with multi-slot frames: upper bound and user multiplexing algorithms," *IEEE Transactions Wireless Communications*, vol. 7, no. 1, January 2008
- [44] N. Ruangchaijatupon and Y. Ji, "Proportional fairness with minimum rate guarantee scheduling in a multiuser OFDMA wireless network," *ACM IWCMC*, Leipzig, Germany, 2009
- [45] F. P. Kelly, A. K. Maulloo, and D.K.H. Tan., "Rate control in communication networks: shadow prices, proportional fairness and stability," *J. of the Operational Research Society*, vol.49, pp. 237-252, April 1998.
- [46] T.-Y. Tsai and Z. Tsai, "Design of a packet scheduling scheme for downlink channel in IEEE 802.16 BWA systems," *IEEE WCNC 2008*.
- [47] IEEE 802.16e-2005, "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands," Dec. 7, 2005.

Reliable Data Forwarding in Wireless Sensor Networks: Delay and Energy Trade Off

M. K. Chahine¹, C. Taddia² and G. Mazzini³

*¹Electronics and Communications Department,
Mechanical and Electrical Engineering Faculty, University of Damascus*

^{2,3}Lepida S.p.A., Bologna

¹Syria

^{2,3}Italy

1. Introduction

Wireless sensor networks (WSNs) are currently the topic of intense academic and industrial studies. Research is mainly devoted to the exploitation of energy saving techniques, able to prolong as much as possible the lifetime of these networks composed of hundreds of battery driven devices[1] [2].

Many envisioned applications for wireless sensor networks require immediate and guaranteed actions; think for example of medical emergency alarm, fire alarm detection, intrusion detection [3]. In such environments data has to be transported in a reliable way and in time through the sensor network towards the sink, a base station that allows the end user to access the data. Thus, besides the energy consumption, that still remains of crucial importance, other metrics such as delay and data reliability become very relevant for the proper functioning of the network [4].

These reasons have led us to investigate a very interesting trade off between the delay required to reliably deliver the data inside a WSN to the sink and the energy consumption necessary to the achievement of this goal.

Typically WSNs consist of many sensor nodes scattered throughout an area of interest that monitor some physical attributes; local information gathered by these nodes has to be forwarded to a sink. Direct communication between any node and the sink could be subject only to just a small delay, if the distance between the source and the destination is short, but it suffers an important energy wasting when the distance increases. Therefore often multihop short range communications through other sensor nodes, acting as intermediate relays, are preferred in order to reduce the energy consumption in the network [5]. In such a scenario it is necessary to define efficient techniques that can ensure reliable communications with very tight delay constraint. In this work we focus our attention on the control of data transport delay and reliability in multihop scenario.

Reliable communications can be achieved thanks to error control strategies: typically the most applied techniques are forward error correction (FEC), automatic repeat request (ARQ) and hybrid FEC-ARQ solutions. A simple implementation of an ARQ is represented by the Stop and Wait technique, that consists in waiting the acknowledgment of each transmitted

packet before transmitting the next one, and retransmit the same packet in case it is lost or wrongly received by the destination. The corrupted data can be retransmitted by the source (non cooperative ARQ). Otherwise data retransmissions may be performed by a neighboring node that has successfully overheard the source data transmission (cooperative ARQ) [4].

We have analyzed, in a previous work [6], four reliable data forwarding methods, based on hybrid FEC and non cooperative ARQ techniques, by focusing the attention mainly on their energy consumption. In particular we have compared the direct and multihop communications by defining the regions in which one is more energy efficient than the other, to ensure a predefined reliability of the communication. Furthermore, in case of multihop path, we have defined regions in which the exploitation of FEC hop-by-hop (detect-and-forward solution) can be helpful and energetic efficient with respect to the use of FEC only at the destination (amplify-and-forward solution).

We extend here this analysis by introducing the investigation of the delay required by the reliable data delivery task. To this aim we investigate the delay required by a cooperative ARQ mechanism to correctly deliver a packet through a multihop linear path from a source sensor node to the sink. In particular we analyze the relation between the delay and the coverage range of the nodes in the path, therefore the relation between the delay and the number of cooperative relays included in the forwarding process. This allows to study optimal multihop topologies to improve data forwarding performance in sensor networks while saving energy as much as possible. The cooperative approach is also compared with other non cooperative solutions, and the delay reduction that the cooperative technique allows to obtain with respect to the more trivial non cooperative ones, is shown. We present analytical expressions for the investigated delay in many scenario and we validate them by means of simulation.

Finally a simple simulation analysis of the energy required by the investigated ARQ techniques has been performed, in order to understand the actual trade off shown by the two approaches.

The rest of the work is organized as follows: Section 2 describes the network topology and the ARQ protocols that we have analyzed; Section 3 provides a general mathematical framework to evaluate the average delay required by the proposed ARQ techniques to deliver a correct packet to the sink and closed equations of the delay in some particular topologies; Section 4 introduces a framework to model the energy consumption involved during the data delivery; Section 5 compares the mathematical model results with those obtained with simulations and shows the delays and the energy consumption of different ARQ techniques; Section 6 concludes the chapter.

2. System model

Consider a multihop linear path composed by a source node (node $n = 1$), a sink (node $n = N$) and $N - 2$ intermediate relay nodes (nodes $n = 2, \dots, N - 1$), equally spaced, as shown in Figure 1. The total path is consequently composed by $H = N - 1$ subsequent links. Suppose that all the nodes have a circular radio coverage and all the nodes in the path have the same transmission range. Let R be the transmission range of each node, expressed in terms of number of links. This means that whenever a node transmits a packet, due to the broadcast nature of the wireless channel, the packet can be received by a set SR of nodes,

composed by all the nodes inside the coverage area of the sender that are in a listen state (consider that most of the Media Access Control (MAC) protocols for WSNs are low duty cycle protocols that awake nodes only when necessary, by letting nodes in a sleep state during the rest of the time to save energy [7]).

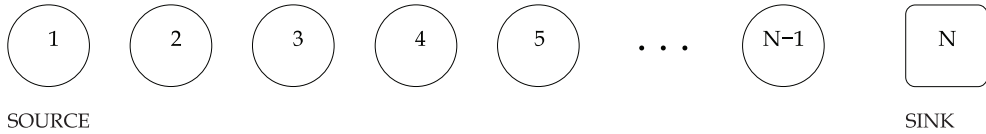


Fig. 1. Linear multihop path between the source node and the destination sink.

For example, by considering $R = 2$ and by referring to Figure 1, when node 3 broadcasts a packet, the packet can be received by the set SR of nodes, with $SR = \{1, 2, 4, 5\}$. Among the set SR we define the subset SF of the possible forwarders, i.e., the nodes that could forward the data towards the destination. By following the strategy suggested by many geographic routing protocols proposed in literature [9], this subset SF includes only the nodes belonging to SR that have a distance to the destination that is lower than the distance between the transmitter and the destination. By referring to the previous example, SF is composed by nodes 4 and 5. Generally, for each node $n \in [1, N - 1]$ that is transmitting a packet, we can define a set SF_n of possible forwarders.

2.1 Cooperative ARQ

The cooperative ARQ strategy allows to exploit the collaboration of more relays overhearing the packet transmitted by a node. This approach supposes that for each node n , all the nodes belonging to the set SF_n are awake and available for the packet reception; the case in which none of them is available will be included as a possible reason of link error packet delivery, as explained in the following mathematical framework (Section 3).

When a node n transmits a packet, the packet is forwarded by the node n_f belonging to the set SF_n , that has correctly received the packet and which is the closest to the destination, in order to complete the data delivery with the minimum number of hops and in the fastest way. Only if no one among the possible forwarder nodes has correctly received the packet, a packet retransmission is requested to the node n ; otherwise the other nodes of SF_n can help the data forwarding process by transmitting the packet, in case they have received it correctly.

Consider, for example, the linear path in Figure 1. The packet delivery process begins from the source node $n = 1$, that broadcasts a packet with range $R = 2$. In this case the forwarder set is $SF = \{2, 3\}$; among these nodes the closest to the destination is $n = 3$. If node $n = 3$ correctly receives the packet it rebroadcasts it; otherwise if it detects that the received packet is not correct the data delivery will continue from node $n = 2$, in case $n = 2$ has correctly received the packet; otherwise the process will begin again from the node $n = 1$ that proceeds by retransmitting the same packet. This procedure is repeated for all the nodes in the path until a correct packet reaches the destination $n = N$.

2.2 Non cooperative ARQ

The non cooperative ARQ strategy defines a transmission range R and schedules communications only between nodes that are R links distant. This means that when a node n

transmits a packet, all the nodes of SF_n , except the node distant R link away, can remain in a sleep state, as they do not need to receive the packet, since they will not be involved in the packet forwarding process. In case the packet has not be correctly delivered to the node $n + R$ a retransmission is requested to the sender node n .

This ARQ strategy is a generalization of the simple hop-by-hop detect-and-forward technique analyzed in [6], where data packet delivery goes on hop-by-hop baiss and possible retransmissions are required to the previous node of the path; clearly the hop-by hop detect-and-forward case can be derived from the general non cooperative ARQ strategy by choosing $R = 1$.

3. Delay: mathematical framework

To evaluate the performance of the ARQ strategies discussed above, we define some performance metrics. We are interested in the delay of the packet delivery process, from the source node to the sink, and in the probability distribution of completing the packet delivery in a certain number of steps (k steps), thus within a certain delay.

By considering that each transmission involves a time slot unit we can proceed by evaluating the delay as multiple unitary time slots and we can calculate it as the number of transmissions needed to deliver a correct packet to the destination. We neglect the delay of ACK or NACK packets. Furthermore when considering wireless communications implicit acknowledgement can also be used [10]: in a multi-hop wireless channel if a node transmits a packet and hears its next-hop neighbor forwarding it, it is an implicit acknowledgement that the packet has been successfully received by its neighbor. The following Subsections (3.1, 3.2, 3.3) present the Markov chains describing the packet forwarding process and the mathematical framework that calculates the average delay and the delay probability distributions for both the cooperative and non cooperative ARQ strategies.

The validity of this mathematical framework has been verified in the previous work [12] by showing a perfect matching between results obtained by means of simulations with the ones obtained by following the mathematic equations given below.

3.1 Transition probabilities

3.1.1 Cooperative ARQ

Let q be the probability to successfully deliver a packet to a node inside the transmitter coverage area; q defines the single transmission success probability between two nodes. So $p = 1 - q$ will be the single transmission error packet probability. For the sake of simplicity the probability q is supposed to be the same inside the coverage area, irrespectively of the distance between the sender and the receiver, provided that they both belong to the subset SF of the sender node. This allows to consider the link error probability not only as a function of the received signal strength, but also dependent on other factors like for example: possible collisions or nodes that are not awake during the packet delivery.

For each node n , the probability to correctly deliver a packet to a node that is R links distant (node $n + R$) is equal to q . So the probability that the packet is not correctly received by this node is $(1 - q)$, while it is correctly received from the immediately previous node ($n + R - 1$) with a probability q . So with a probability $(1 - q)q$ the packet will be forwarded by the node $n + R - 1$. If also this node has not correctly received the packet sent by node n , event that occurs with a probability $(1 - q)^2$, with a probability $(1 - q)^2q$ the packet will be

forwarded by the node $n + R - 2$. If none of the nodes between node $n + 1$ and node $n + R$ receives a correct packet it is necessary to ask the retransmission of the packet by the node n . It is possible to describe the process concerning one data packet forwarding from the source node $n = 1$ to the destination $n = N$ with a discrete time Markov chain. We identify each node in the path with a number n , where n varies from 1 (the source) to N (the destination). Each state in the chain represents a node in the path: in particular the process is in state n at a certain time when n is the furthest node, starting from the source, that has correctly received a packet until that time and it has to carry on the forwarding process.

We define $P_{n,n+j}$ as the transition probability between a state n and the state $n + j$. $P_{n,n+j}$ represents the probability that the data packet broadcasted by node n has been correctly received by node $n + j$ while it has not been correctly received by the other nodes belonging to the subset SF_n that are closer to the destination N with respect to the node $n + j$; in other words, $P_{n,n+j}$ is the probability that the next forwarder will be node $n + j$, given that the transmitting node was node n . $P_{n,n+j}$ can be calculated as follows:

- if $1 \leq n \leq N - R$:

$$P_{n,n+j} = q(1 - q)^{R-1} \quad \text{if } 1 \leq j \leq R$$

$$P_{n,n+j} = (1 - q)^R \quad \text{if } j = 0$$

$$P_{n,n+j} = 0 \quad \text{otherwise}$$

- if $N - R + 1 \leq n \leq N - 1$:

$$P_{n,n+j} = q(1 - q)^{N-n-j} \quad \text{if } 1 \leq j \leq N-n$$

$$P_{n,n+j} = (1 - q)^{N-n} \quad \text{if } j = 0$$

$$P_{n,n+j} = 0 \quad \text{otherwise}$$

- if $n = N$:

$$P_{n,n+j} = 1 \quad \text{if } j = 0$$

$$P_{n,n+j} = 0 \quad \text{otherwise}$$

Note that there are different $P_{n,n+j}$ equations depending on which state n we are considering. For nodes n , with $1 \leq n \leq N - R$, the transition probability from node n to node $n + j$, with $1 \leq j \leq R$, is equal to $q(1 - q)^{R-j}$. In fact, it takes into account that the maximum distance that is possible to cover during a transmission is equal to R links; so if the packet is correctly detected by node $n + R$ we have the transition probability between state n and state $n + R$, with a transition probability $P_{n,n+R} = q$; in case that $i = R - j$ nodes do not correctly receive the packet, there is a transition between state n and state $n + j$, with probability $P_{n,n+j} = q(1 - q)^{R-j}$; j can vary between 1 and R , representing the number of relays belonging to the subset SF_n . The last $R - 1$ nodes that precede the destination node (nodes n with $N - R + 1 \leq n \leq N - 1$) represent an exception, since the distance between the transmitting node and the destination is less than the transmission range of the nodes and therefore in their subsets SF there are less possible cooperative relay nodes.

An example of Markov chain for a path composed by four nodes ($N = 4$), $H = N - 1 = 3$ links and range $R = 2$ is shown in Figure 2, for which we write the transition probability matrix P_C as a function of the success link probability.

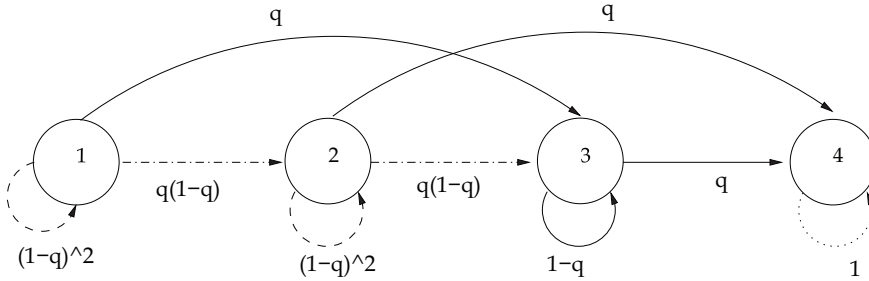


Fig. 2. Markov chain for the topology $N = 4$, $H = 3$, $R = 2$.

$$P_C = \begin{pmatrix} (1-q)^2 & (1-q)q & q & 0 \\ 0 & (1-q)^2 & (1-q)q & q \\ 0 & 0 & 1-q & q \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (1)$$

The same matrix P_C expressed as a function of the error link probability becomes:

$$P_C = \begin{pmatrix} p^2 & (1-p)p & 1-p & 0 \\ 0 & p^2 & (1-p)p & 1-p \\ 0 & 0 & p & 1-p \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (2)$$

A similar approach was used in [8] to evaluate the mean number of hops required to realize the Route Request Process by the Ad hoc On-Demand Distance Vector (AODV) routing for ad hoc networks. The approach used here is quite different since it takes into account all the possible retransmissions of the wrong packets.

Note that the Markov chain is characterized by $N - 1$ transient states (the source node $n = 1$ and all the other relays $n = 2, 3, \dots, N - 1$) and by an absorbing state (the destination sink, node $n = N$, characterized by a transition probability $P_{N,N} = 1$). In fact a state n of a Markov chain is defined as transient if a state i , with $i \neq n$, exists that is accessible from state n while n is not accessible from i ; once the system is in state n it can go into one of the states $i = n + j$, with $j \leq \min\{R, N - n\}$ but once the system is in this state $n + j$ it means that the packet has arrived correctly, at least at node $n + j$ therefore node n will not need to retransmit it again; so state $n + j$ is accessible from state n and state n is not accessible from state $n + j$. State N as an absorbing state is a good representation of the physical process that we are analyzing: in fact, this Markov chain describes the packet forwarding process, the travel of a packet from a source towards a destination, where the packet stops and does not have to go in any other place. Results obtained by simulations and presented in the following Section will confirm the correctness of this model.

3.1.2 Non Cooperative ARQ

In case of the non cooperative ARQ the process is composed by a total number of states equal to the ratio $\lceil \frac{H}{R} \rceil + 1$. In fact, as Figure 3 shows, after choosing the range R there are some nodes that will never be involved in the packet forwarding process: for example node 2 in Figure 3 when $R = 2$. For each state n of the chain there is a probability $1 - p$ that at the next step the packet will be forwarded by the next state of the chain (node $n + \min\{R, N - n\}$) and a probability p that it will be retransmitted by the node n .

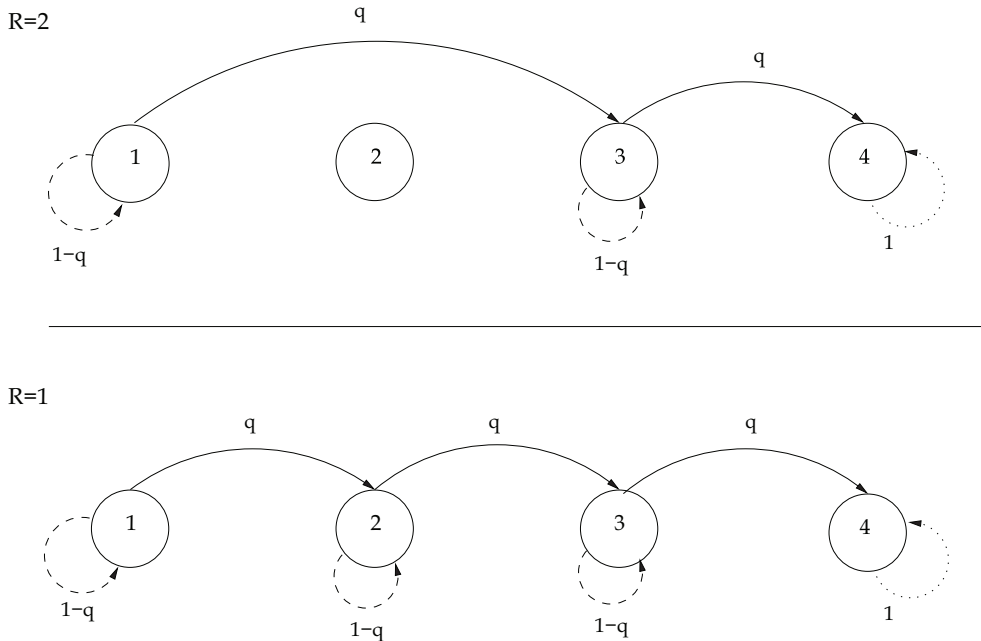


Fig. 3. Markov chain for the topology $N = 4, H = 3$. Non cooperative ARQ with $R = 2$ in the top of the Figure and with $R = 1$ in the bottom of the Figure.

The transition probability matrix is a matrix of dimension $(\lceil \frac{H}{R} \rceil + 1) \times (\lceil \frac{H}{R} \rceil + 1)$:

$$P_{NC} = \begin{pmatrix} p & 1-p & 0 & 0 & \dots & 0 \\ 0 & p & 1-p & 0 & \dots & 0 \\ 0 & 0 & p & 1-p & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \tag{3}$$

3.2 Delay probabilities distribution

For a generic state i of a discrete time Markov chain [11] described by a generic matrix P of transition probabilities, we define the time of first visit into state i as: $T_i = \inf \{k \geq 1 \mid X_k = i\}$,

where k is the number of visits into the state i and X_k is the state in which the system is at time k . Generally we denote by $f_{i,j}^{(k)}$ the probability that a system described by a discrete time Markov chain transits for the first time from state i to state j in k steps. This probability is defined as: $f_{i,j}^{(k)} = P\{T_j = k | X_0 = i\}$, where X_0 is the initial state of the system. Chapman-Kolmogorov equations states that the probability $f_{i,j}^{(k)}$ can be calculated as a sum of all the possible combinations of the probabilities of going from state i to state j by going, during the intermediate steps, through the other states of the systems, apart from the state j , that has to be reached for the first time at the step k . Formally we have:

$$f_{i,j}^{(k)} = \sum_{s_1, s_2, \dots, s_{k-1} \in S \setminus \{j\}} P_{is_1} \cdot P_{s_1 s_2} \cdot \dots \cdot P_{s_{k-1} j} \quad (4)$$

where S is the total space of the states and $P_{S_i S_y}$, (with $i, y \in 1, \dots, k-1$), are the transitions probabilities of the matrix P . For each $k \geq 1$ this can be written also as: $f_{i,j}^{(k)} = P_{i,j}^{(k)} - \sum_{i=1}^{k-1} f_{i,j}^{(i)} P_{i,j}^{(k-i)}$. This suggests to calculate the $f_{i,j}^{(k)}$ in a recursive way through the knowledge of the transition probabilities included in the matrix P . For a finite state Markov chain, Equation 4 can be represented in a matrix form: $f_{i,j}^{(k)}$ results to be the element in position (i, j) of the $k - th$ power of the matrix \tilde{P} , where \tilde{P} is equal to matrix P except for the $j - th$ row that is taken as a null row in order to remove the possibility of passing through the $j - th$ state in an intermediate step $k' < k$.

3.2.1 Cooperative ARQ

According to the general definitions given above, we can derive the delay probability distribution in the specific case of the Markov chain described by the matrix P_C . The probability distribution of ending the process in a certain number k of steps is expressed by the probability that the system transits for the first time from state 1 to state N after k steps. The number of visits for each transient state varies accordingly to the link error probability and to the probability that no one of the relays belonging to the subset SF_n of a node n correctly receive the packet and therefore needs to ask for a retransmission of the packet to the sender node n . The number of visits to state N is infinite: once the packet arrives at destination the process is ended, it remains into the absorbing state for an infinite time. In fact, in the long term behavior, when time tends to infinity, the steady state probability of state N is one while for all the other transient states n we have $\lim_{k \rightarrow \infty} P_{C,i,n}^{(k)} = 0, \forall i$, i.e., each state will be absorbed into state N . The delay that we are going to evaluate is therefore the mean time of the first visit to state N . The Markov chain in fact refers to the delivery of a single packet from the source towards the destination; when considering the transmission of another packet from the source node the process begins again from the state 1 of the Markov chain.

We indicate the probability that the packet is correctly forwarded to the destination in a number of steps k for the cooperative ARQ is defined as:

$$f_{C,1,N}^{(k)} = \sum_{s_1, s_2, \dots, s_{k-1} \in S \setminus \{N\}} P_{1s_1} \cdot P_{s_1 s_2} \cdot \dots \cdot P_{s_{k-1} N} \quad (5)$$

This can be easily calculated as the element in position $(1, N)$ of the $k - th$ power of the matrix \tilde{P}_C , where \tilde{P}_C is built equal to matrix P_C except for the element (N, N) that is 0

instead of 1. These probabilities are a function of the number of hops H composing the path and the range R , so it is useful to indicate this dependency by calling these probabilities in the rest of the chapter as $f_{C_{1,N}}^{(k)}(R, H)$. We have found out that for some particular values of the transmission range ($R = 1$ and $R = H$) the probability can be expressed through simple closed form equations. So we have $f_{C_{1,N}}^{(k)}(1, H) = \binom{k-1}{H-1} p^{k-H} (1-p)^H$ and $f_{C_{1,N}}^{(k)}(H, H) = p^k (1-p)$.

3.2.2 Non cooperative ARQ

The probability $f_{C_{1,N}}^{(k)}(R, H)$ can be calculated by following the general approach described at the beginning of subsection 3.2 applied to the matrix P_{NC} . Note that $f_{C_{1,N}}^{(k)}(R, H)$ results to be described by the following closed equation:

$$f_{NC_{1,N}}^{(k)}(R, H) = \binom{k-1}{\lceil \frac{H}{R} \rceil - 1} p^{k - \lceil \frac{H}{R} \rceil} (1-p)^{\lceil \frac{H}{R} \rceil} \tag{6}$$

3.3 Average delay

The average delay is represented by the absorption time into last state of the chain starting from the source. The mean time of first visit from state i to state j of a discrete time Markov chain, called $T_{i,j}$ is defined as follows:

$$T_{i,j} = \begin{cases} \infty & \text{if } \sum_{k=1}^{\infty} k f_{i,j}^{(k)} < 1 \\ \sum_{k=1}^{\infty} k f_{i,j}^{(k)} & \text{if } \sum_{k=1}^{\infty} k f_{i,j}^{(k)} = 1 \end{cases}$$

When $\sum_{k=1}^{\infty} f_{i,j}^{(k)} = 1$ the time $T_{i,j}$ is univocally solution of the following equation:

$$T_{i,j} = 1 + \sum_{s \neq j} P_{i,s} T_{s,j} \tag{7}$$

By fixing an arrival state j , equation 7 allows to obtain a linear system whose solutions are the mean time of first transition from each one of the possible initial states i , ($\forall i \in S \setminus \{j\}$, where S is the total space of the states), to the final state j .

3.3.1 Cooperative ARQ

According to the general definitions given above, we can derive the average delay in the specific case of the Markov chain described by the matrix P_C . The delay we want to evaluate is the absorption time to state N by starting from state 1, i.e., the mean time of first visit from state 1 to state N . Since in our case the state N is an absorbing state the condition $\sum_{k=1}^{\infty} f_{1,N}^{(k)} = 1$ is verified; in fact the probability for each transient state to be absorbed into

state N is equal to one. So we can calculate the mean time $T_{C_{i,N}}$ by solving the linear system defined in Equation 7, where the transition probabilities $P_{i,s}$ are taken from the matrix P_C :

$$T_{C_{i,N}} = 1 + \sum_{s \neq N} P_{i,s} T_{C_{s,N}}$$

where $i = 1, 2, \dots, N - 1$.

Since it is a function of the number of links H composing the path and of the range R , in the rest of the chapter the term $T_C(R,H)$ refers to that quantity. We omit the indexes 1, N defining the starting and the final node, for the sake of simplicity, since they nevertheless are always the source node 1 and the destination N . We have analyzed the possibility to express the delay in a closed form, for each value of the total number of links composing the path, H , and for some particular values of the transmission range: $R = 1$, $R = H$, $R = H - 1$ and $R = H - 2$. When $R = 1$ the delay has the following expression:

$$T_C(1,H) = H / (1 - p) \quad (9)$$

When $R = H$ we have:

$$T_C(H,H) = 1 / (1 - p) \quad (10)$$

When $R = H - 1$ we have found the following Equation:

$$T_C(H-1,H) = \frac{2 + \sum_{i=1}^{H-2} p^i}{(1-p) \sum_{i=1}^{H-2} p^i} \quad (11)$$

$$= \frac{1}{1-p} + \frac{p}{p-p^H} \quad (12)$$

When $R = H - 2$ the following expression is valid:

$$T_C(H-2,H) = \frac{1}{(1-p) [\sum_{i=1}^{H-3} p^i]^2} \quad (13)$$

$$\cdot \left[\sum_{i=1}^{H-5} (3+i)p^{i+1} + \right. \quad (14)$$

$$\left. + Hp^{H-3} + \sum_{i=0}^{H-4} p^{H-2+i}(H-3-i) \right] \quad (15)$$

$$(16)$$

$$= \frac{p^4(2-p) + p^{2H} + p^{H+1}[1-5p+2p^2]}{(1-p)[p^2 - p^H]^2} \quad (17)$$

3.3.2 Non cooperative ARQ

The average delay required by the non cooperative approach can be derived by following the general approach described above and applied to the matrix P_{NC} . It can also be derived by simply thinking that is the product between the mean number of hops in which the total path is divided once the transmission range R has been chosen, (that turns to be $\lceil \frac{H}{R} \rceil$), and the mean number of transmission needed to correctly deliver a packet between two nodes R links distant. Suppose that p is the link error probability at distance R . We call P_a the probability to make a attempts in order to deliver a correct packet in a single hop communication; P_a can be calculated as the probability to make one successful transmission (event that happen with a probability $1 - p$) and $a - 1$ failures (event happening with probability p^{a-1}). The mean number of transmissions $E[tx]$ needed per single hop is derived as follows:

$$E[tx] = \sum_{a=1}^{\infty} aP_a \quad (18)$$

$$= \sum_{a=1}^{\infty} ap^{a-1}(1-p) = \frac{1}{1-p} \quad (19)$$

The delay is therefore calculated as:

$$T_{NC}(R, H) = \frac{\lceil \frac{H}{R} \rceil}{1-p} \quad (20)$$

4. Energy model

In order to better evaluate the performance of the proposed ARQ strategies, we also investigate the energy consumption required by them in different scenarios. This allows to obtain useful trade offs between energy consumption and delay requested to accomplish a task.

We define a simple energetic model, by referring to the considerations made in [6]. Suppose having a scenario with H hops between the source and the destination and having fixed distance between two subsequent nodes.

In more detail, the energy E required in a point-to-point communication between two nodes is the sum of two contributions: the energy spent by the transmitter for transmitting a packet, E_{TX} , and the energy spent by the node receiving the data packet, indicated with E_{RX} . More in detail the energy $E_{TX} = E_c + E_d(R)$ comprises two contributes: the energy spent by the transceiver electronics and by the processor to encode the packet with a preselected FEC code to reveal the errors in the packet, E_c and a contribution $E_d(R)$ proportional to the distance between the nodes involved in the communication and the signal to noise ratio desired at the destination. The energy E_{RX} comprises the energy of the transceiver electronics and the energy spent by the processor in decoding the packet, E_c . The total energy required to deliver a correct packet to the destination, E_{TOT} , can be calculated as the energy spent for a transmission multiplied by the total number of transmissions performed during the

forwarding process, N_{TX} , added to the energy spent for a reception multiplied by N_{TR} , the total number of receptions occurred during the forwarding process: $E_{TOT} = N_{TX}E_{TX} + N_{RX}E_{RX}$. Let α be the ratio between E_c and the term $E_d(1)$, that is the contribution of energy E_d required to send a packet to a node that is 1 link distant from the sender: $\alpha = E_c/E_d(1)$. We proceed by normalizing the total energy with respect to the contribute $E_d(1)$. Therefore the normalized energy $\hat{E}_{TOT}(R, H)$ for a path composed by H links and with a transmission range R is:

$$\hat{E}_{TOT}(R, H) = (R^\eta + \alpha)N_{TX} + \alpha N_{RX} \quad (21)$$

where N_{TX} and N_{RX} refers to the specific total number of transmissions and receptions of the ARQ strategy under analysis and η is the path loss exponent.

5. Numerical results: delay-energy trade off

Results related to the performance in terms of delay and energy consumption of the two mentioned ARQ approaches and their correlations and dependencies with various parameters, such as the communication range R and the sensor node circuitry (with the parameter α) has been deeply investigated and presented in the previous work [12].

In this Section we rather show the performance of the proposed cooperative and non cooperative ARQ strategies in terms of delay and energy consumption, by pointing up the trade off between these two important metrics.

In order to monitor also the comparison between the two ARQ approaches, we investigate in our trade off study the ratio between the results obtained with the cooperative solution and the non cooperative one, for both two metrics, delay and energy consumption.

The results presented in this section have been tested by means of simulations by following the energetic model described in Section 4. Let us precise that the results presented in the following have been obtained only by means of simulation, since it is not trivial to derive a precise mathematical model to calculate the number N_{RX} for the cooperative ARQ. In fact the number of nodes receiving the packet or each packet transmissions depends on the node that is transmitting: by referring to the matrix P_C , it depends on the state n of the sender node: the number of receiving nodes for each packet transmission is R if $1 \leq n \leq N - R$, but it is less than R for the states n of the chain that are $N - R + 1 \leq n \leq N - 1$.

Figures 4 and 5 shows the tradeoff between the delay and the energy consumption. As an example a path composed by $H = 10$ hops has been considered. The communication range of the nodes has been taken equal to $R = 3$ or $R = 5$ and different values of the parameter $\alpha = 5, 15, 30$ has been tested. In order to compare the different ARQ mechanisms in a realistic scenario, we have estimated the range of values of the parameter α by referring to an actual sensor node, the μ AMPS1, as followed in [6]. We observe that for these specific hardware constraints the parameter α can vary in a range between 1 and 50. We have used values of α between these boundaries to compare the energy spent by the different ARQ strategies. Accordingly to the scenario parameters (R, α) and as a function of the channel quality P this graph allows to easy calculate the gain achievable in terms of energy and latency by choosing one of the two proposed ARQ approaches.

Accordingly to the scenario parameters (R, α) and as a function of the channel quality P this graph allows to easy calculate the gain achievable in terms of energy and latency by choosing one of the two proposed ARQ approaches.

Figure 4 shows as x-axis the ratio between the delay of the cooperative ARQ technique and the delay of the non cooperative one and as y-axis the ratio between the energy consumption required by the cooperative approach and the non cooperative one. In this graph the cooperative and non cooperative techniques have been implemented with the same communication range for each node. While in Figure 5 the comparison concerns the cooperative ARQ with a generic range R and the non cooperative solution implemented with communication range $R = 1$ (hop-by-hop detect-and-forward case). In both the Figures, results are plotted for different values of the link error probability p , varying between 0.1 and 0.9, as indicated in the graphs.

Figure 4 evidences that while the delay required by the cooperative solution is always less than the non cooperative one, a trade off is present concerning the energy consumption, that

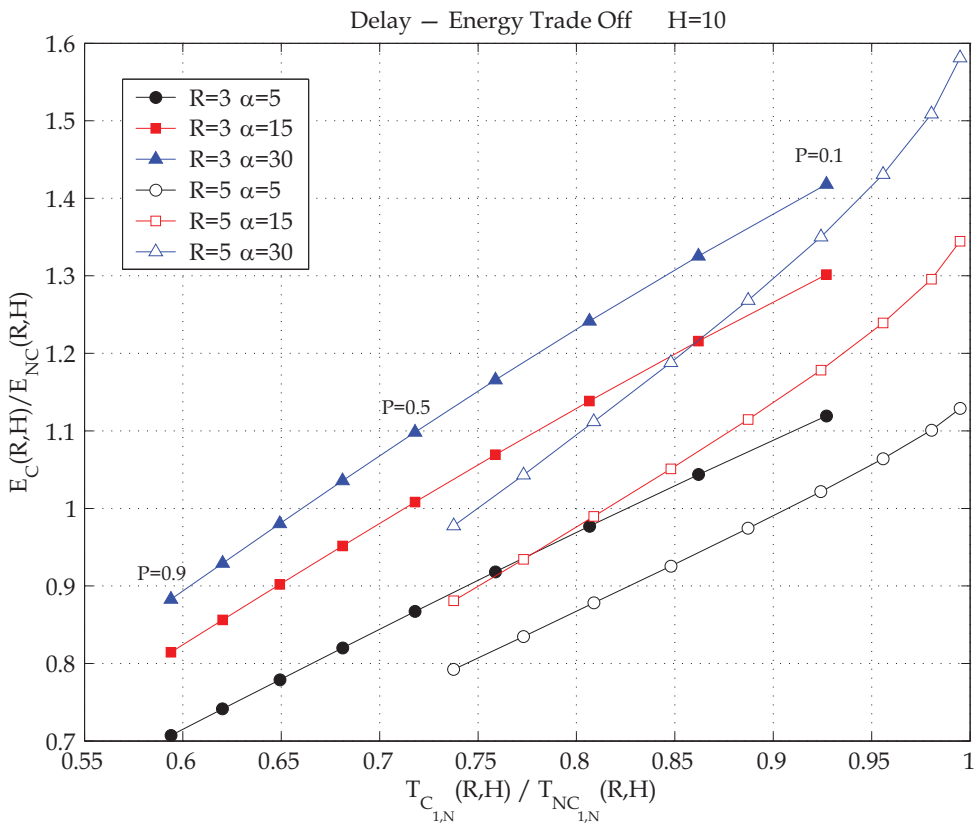


Fig. 4. Delay-Energy tradeoff. Comparison between the cooperative and the non cooperative ARQ techniques both with the same communication range R for the nodes. The path is composed by $H = 10$ links.

depends on the ratio α , on the packet error probability per link p and on the range R . In particular, we can see that the cooperative ARQ turns out to be an energetic efficient solution with respect to the non cooperative ARQ when the link reliability is quite low and when the ratio α is sufficiently low. Performance in terms of delay reduction are even bigger if comparing the cooperative ARQ (with range R) with the non cooperative single hop detect- and-forward ($R = 1$), as evidenced in Figure 5. Also in this case a trade off between delay and energy can be achieved: notice that there are regions of p and α (when α is sufficiently low in this case) for which the cooperative ARQ, besides giving better delay performance, also can help in saving the nodes energy and thus extending the network lifetime.

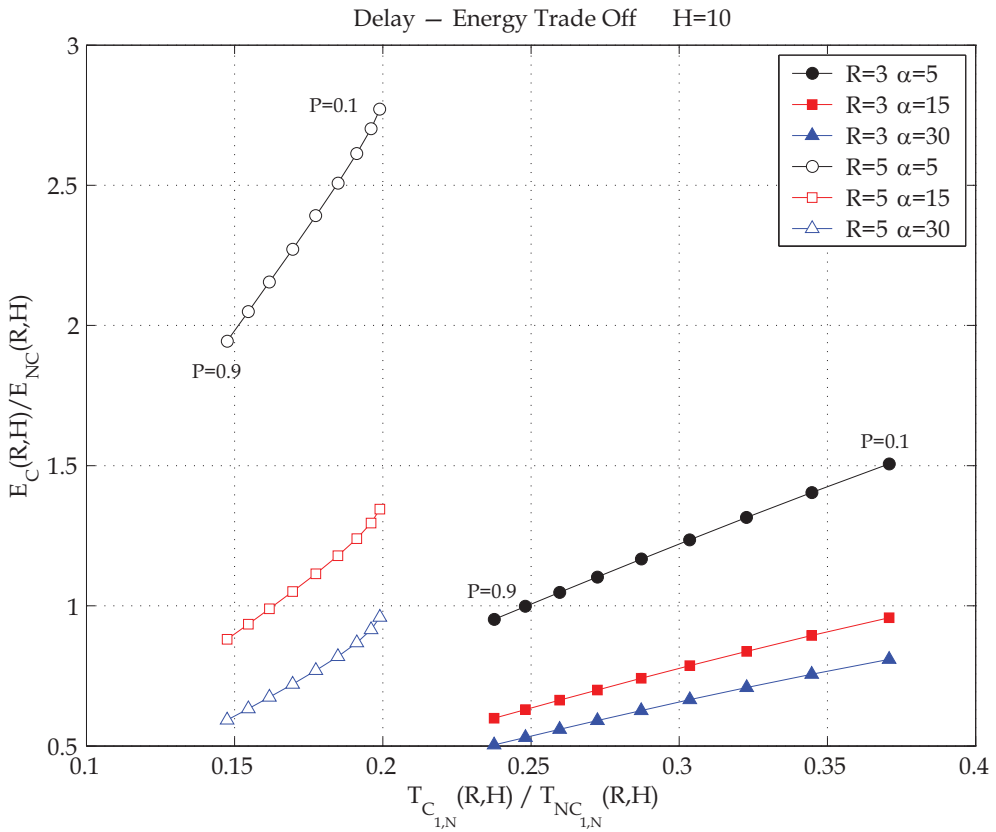


Fig. 5. Delay-Energy tradeoff. Comparison between the cooperative ARQ technique with communication range R and the non cooperative ARQ single hop decode-and-forward approach (with $R = 1$). The path is composed by $H = 10$ links.

6. Conclusions

This work has deeply presented an important trade off between energy consumption and delay in the task of reliable data delivery between a source node and a destination sink in a wireless sensor network. We have presented the performance in terms of delay and energy consumption of cooperative and non cooperative ARQ techniques that allows to ensure reliable communications in WSNs for delay constraints applications. Our investigations have showed that the proposed cooperative ARQ is a successful technique. In particular the cooperative solution, besides showing always better performance concerning the timeliness of data delivery, with respect to the non cooperative approach, can in some scenario outperform the trivial non cooperative hop-by-hop detect and forward technique also in terms of energy saving.

7. References

- [1] W.Ye, J. Heidemann, D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks", *Infocom 2002*, 23-27 June, pp. 1567 - 1576, vol.3.
- [2] V. Raghunathan, C. Schurgers, Sung Park, M. Srivastava, "Energy-aware wireless microsensor network", *IEEE Signal Processing Magazine*, March 2002, pp.40-40, vol.19.
- [3] L. Bernardo, R. Oliveira, R. Tiago, P. Pinto, "A Fire Monitoring Application For Scattered Wireless Sensor Networks". *WinSys 2007*, 28-31 July, Barcelona.
- [4] I. Cerutti, A. Fumagalli, P.Gupta, "Delay Models of Single-Source Single-Relay Cooperative ARQ Protocols in Slotted Radio Networks with Poisson Frame Arrivals", *Infocom 2007*, pp. 2276-2280, vol.16
- [5] Z. Shelby, C. Pomalaza-Raez, J. Haapola, "Energy optimization in multihop wireless embedded and sensor networks", *PIMRC 2004*, 5-8 Sept, pp. 221-225, vol.1.
- [6] C. Taddia, G. Mazzini, "On the Energy Impact of Four Information Delivery Methods in Wireless Sensor Networks", *IEEE Communication Letters*, Feb. 2005, Vol. 9, n. 2, pp. 118-120.
- [7] S. Ramakrishnan, H. Huang, M. Balakrishnan, J. Mullen, "Impact of sleep in a wireless sensor MAC protocol", *VTC Fall 2004*, 26-29 Sept, pp.4621-4624, vol. 7.
- [8] C.Taddia, G.Mazzini, "An Analytical Model of the Route Acquisition Process in AODV Protocol", *IEEE WirelessCom 2005*, 13-16 June, Hawaii.
- [9] M. Zorzi, R. Rao, "Geographic random forwarding (GeRaF) for ad hoc and sensor networks: multihop performance", *IEEE Transaction on Mobile Computing*, pp. 337-348, vol.2, issue 4, 2003.
- [10] T. Hwee-Pink, K.G. Winston, L. Doyle, "A Multi-hop ARQ Protocol for Underwater Acoustic Networks", *Proceedings of the IEEE/OES OCEANS Conference*, 18-21 June 2007, Aberdeen, Scotland.
- [11] S. Karlin, H.M. Taylor, "A First Course in Stochastic Processes", Academic Press.

- [12] C. Taddia, G.Mazzini, M.K.Chahine, K. Shahin, "Reliable Data Forwarding for Delay Constraint Wireless Sensor Networks", International Conference on Information and Communication Technologies, ICTTA 2008, 7-11 April, Damascus, Syria.

Cross-Layer Connection Admission Control Policies for Packetized Systems

Wei Sheng and Steven D. Blostein
*Queen's University, Kingston,
ON Canada*

1. Introduction

Delivering quality of service in packetized mobile cellular systems is costly, yet critical. Recently, cross-layer connection admission control policies [1] [2] have been shown to realize network performance objectives for multimedia transmission that include constraints on delay and blocking probability. Current third generation (3G) systems such as high speed uplink packet access (HSUPA) employ a threshold-based admission control (AC) policy to reserve capacity to increase quality of service (QoS). In threshold-based AC, a user request is admitted if the load reported is below a threshold. Although a threshold-based AC policy is simple to implement and may be improved upon to take into account resource allocation information [3], it unfortunately cannot meet upper layer QoS requirements, such as required in the data-link and network layers [4].

In this chapter, AC policies are investigated for packetized code division multiple access (CDMA) systems that can both maximize overall system throughput and simultaneously guarantee quality of service (QoS) requirements in both physical and upper layers. To further improve user capacity, multiple antennas are employed at the base station, and a truncated automatic repeat request (ARQ) scheme is employed in the data link layer of the system under investigation. Truncated ARQ is an error-control protocol which retransmits an erroneous packet until either it is correctly received or until a maximum number of retransmissions is reached.

The design of optimal connection admission control policies for a packetized CDMA system that incorporates an advanced multi-beamformer basestation at the physical layer and ARQ at the data link layer has, to the authors' knowledge, not been addressed previously. For example, the call level admission control policies for CDMA systems in [4] [5] [6] only focus on circuit-switched networks, in which radio resources allocated to a user are unchanged throughout the call connection, leading to inefficient utilization of system resources, especially for bursty multimedia traffic. In [7] [8], the CAC problem is extended to packet-switched CDMA systems. Unfortunately, the CAC modelling in [7] [8] has been limited to optimizing power control and admission control policies to specific systems, in which physical layer performance, characterized in terms of signal-to-interference (SIR) in each service class, is static. With multiple antennas systems, which are widely employed in current 3G CDMA systems [9] - [14], the physical layer performance depends not only on system state, but also on factors such as spatial angle of arrival (AoA). Therefore, the existing CAC framework in [7] [8] cannot adequately incorporate multiple antenna base-

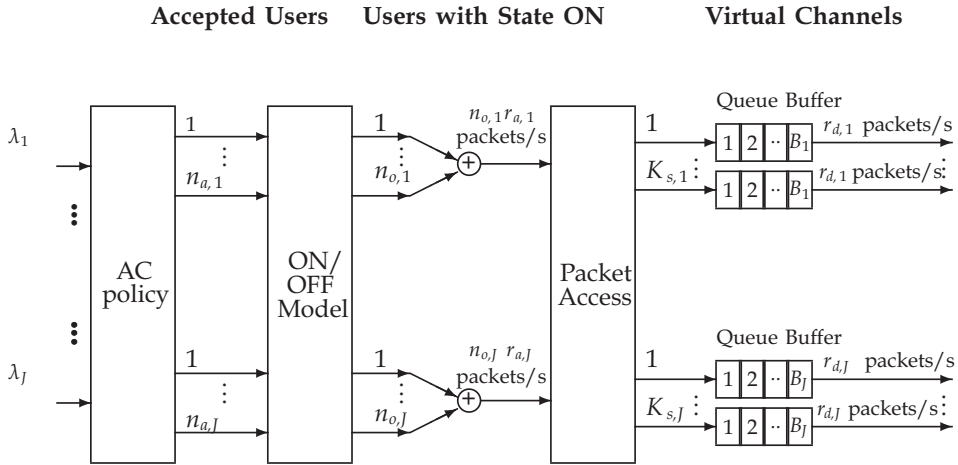


Fig. 1. Signal model for packet-switched networks.

stations. Furthermore, in the above- mentioned design of optimal connection admission control policies, there is no automatic retransmission request (ARQ) mechanism built into the connection admission control design, and therefore is lacking in error control capability.

We remark that in previous work in [15], a packet-level admission control policy is proposed, which dramatically improves system performance by employing both multiple antennas and ARQ. However, the AC scheme is designed at the packet level, in which connection level QoS, such as blocking probability and connection delay, is ignored. Therefore, this packet level AC policy cannot work well for a connection-oriented packet based network. Moreover, AC policies performed at the packet level, instead of at the connection level, may incur implementation difficulties. This fact motivates an investigation into a connection level admission control policy for packet-switched networks with guaranteed QoS constraints at physical, connection and packet levels. In [15], the ARQ and admission control schemes are both performed at the packet level, while in this chapter, the admission control is performed at the connection level, while retransmissions are still performed at packet level, as is widely adopted in practical systems.

The rest of this chapter is organized as follows: the signal model and problem formulation are presented in Sections II and III, respectively. In Sections IV and V, packet-level and physical-layer QoS requirements in terms of packet loss probability and outage probability are analyzed, respectively. An optimal connection admission control policy is derived in Section VI. Numerical results are presented in Section VII.

2. Signal model

A. Traffic model

The signal model is illustrated in Figure 1. We consider an uplink CDMA beamforming system with M antennas at the basestation. A spatial matched filter corresponding to each user in the system is assumed. In addition, suppose there are J classes of statistically independent traffic in the network. The arrival process of the aggregate connections is modeled by a Poisson process with rate λ_j for each class j , where $j = 1, \dots, J$. The duration for each connection is assumed to be exponentially-distributed with mean $\frac{1}{\mu_j}$.

Whenever a connection arrives, the connection admission control (AC) policy, derived offline and implemented as a lookup table, decides whether or not the incoming connection should be accepted. In Figure 1, $n_{a,j}$ denotes the number of accepted users for class j , where $j = 1, \dots, J$. The system state, representing the number of accepted users for each class, is defined as $\mathbf{s} = [n_{a,1}, \dots, n_{a,J}]$. To reduce the size of the state space, no queue buffer is implemented at the connection level, which implies that if the incoming connection is not accepted immediately, it is blocked.

B. Signal model at the packet level

The connection admission control policy decides whether an incoming connection should be accepted. If accepted, a sequence of packets is generated and transmitted over the channel. Following the truncated ARQ protocol, erroneously received packets are retransmitted until correctly received or until a prescribed number of maximum allowed retransmissions is reached.

Continuing along to the right of Figure 1, for each accepted connection, packet-generating traffic is modelled as an ON/OFF Markov process. That is, when a user is in an ON state, packets are generated with a rate $r_{a,j}$ packets per second and when the user is at OFF state, no packets are generated.

For a class j connection, the transition probabilities from ON state to OFF state, or from OFF state to ON state, are denoted by α_j and β_j , respectively. Denote p_{ON}^j as the probability that a

class j user is in the ON state, which can be obtained by $p_{ON}^j = \frac{\beta_j}{\alpha_j + \beta_j}$. Given $n_{a,j}$ accepted users, the number of users in the ON state, denoted by $n_{o,j}$, is a Binomial-distributed random variable. With $n_{o,j}$ users in the ON state, the overall arrival rate for class j is given by $n_{o,j}r_{a,j}$.

In contrast to a circuit-switched network in which each user is allocated a dedicated channel with a fixed transmission data rate, for packet switched networks, no dedicated channels are allocated. Instead, all generated packets from users of a certain class, j , access a given number of shared *virtual channels* denoted by $K_{s,j}$. The value of $K_{s,j}$ is determined by the number of accepted users, the traffic model as well as the QoS requirements. The packets allocated to a class j virtual channel are stored in a packet queue buffer of size B_j , where $j = 1, \dots, J$. The packets in each virtual channel are then transmitted at a rate $r_{d,j}$.

In this chapter, we consider a truncated ARQ scheme (not shown in the figure) which retransmits an erroneous packet until it is successfully received or until the number of maximum allowed retransmissions, denoted by L_j for class j packets, is reached, where $j = 1, \dots, J$. Once a packet is received, the receiver sends back an acknowledgement (ACK) signal to the transmitter. A positive ACK indicates that the packet is correctly received while a negative ACK indicates an incorrect transmission. If a positive ACK is received or the maximum number of re-transmissions, denoted by L_j , is reached, the packet releases the virtual channel and a packet in the queue can then be transmitted. Otherwise the packet will be retransmitted.

C. Signal model at the physical layer

We consider a CDMA beamforming system with an array of M antennas at the base station (BS). At the receiver, a spatial-temporal matched-filter receiver is employed. With

$K = \sum_{j=1}^J K_{s,j}$ virtual channels, there are at most K packets simultaneously transmitted. The received signal-to-noise-plus-interference ratio (SINR) for a desired packet k , where $k = 1, \dots, K$, can be written as

$$\text{SINR}_k = \frac{W}{R_k} \frac{p_k \phi_{kk}^2}{\sum_{i=1, i \neq k}^K \phi_{ik}^2 + \eta_0 W} \quad (1)$$

where W and R_k denote the bandwidth and data rate for the virtual channel allocated to the k -th packet, respectively. The ratio $\frac{W}{R_k}$ represents the processing gain of the CDMA system.

In (1), $p_k = P_k G_k^2$ denotes the received power which is comprised of transmitted power P_k and link gain G_k . The quantity ϕ_{ik}^2 denotes the fraction of packet i 's signal power that passes through the spatial filter (beamformer) corresponding to the spatial response of desired packet k , which can be expressed as $\phi_{ik}^2 = \left| \mathbf{a}_k^H \mathbf{a}_i \right|^2$, in which \mathbf{a}_i denotes the normalized M -dimensional array response vector for packet i , and $(\cdot)^H$ denotes conjugate transpose. The constant η_0 represents the one-sided power spectral density of the background additive white Gaussian noise.

3. Problem formulation

The connection-level and physical-layer QoS can be characterized by blocking probability and outage probability, respectively, while the packet-level QoS can be represented by packet loss probability, defined as the probability that a packet in an accepted connection cannot be delivered to the receiver. Other packet level QoS constraints, such as packet access delay, can be ensured by packet access control, which is not discussed in this chapter.

There exists a performance tradeoff across the different layers. For example, improving connection level performance allows more accepted connections, which leads to an increased aggregate packet generation rate. When the packet generation rate exceeds the packet departure rate, extra packets should be dropped, degrading packet level performance. Although packet level performance can be improved by increasing the number of allocated channels, the physical layer performance degrades with an increased number of channels due to multi-access interference. The proposed cross-layer connection admission control policy should be designed to determine these tradeoffs across different layers.

To characterize overall system performance across different layers, the system throughput, defined here as the number of correctly received packets per second, for a certain admission policy π , can be expressed in terms of the above previously defined quantities as

$$\text{Throughput}(\pi) = \sum_j \lambda_j (1 - P_b^j(\pi)) (1 - P_{out}^{av}(\pi)) P_{ON}^j r_{a,j} (1 - P_L^j(\pi)) (1 - \rho_e^j(\pi)) \quad (2)$$

where $P_b^j(\pi)$, $P_{out}^{av}(\pi)$, $P_L^j(\pi)$ and $\rho_e^j(\pi)$ denote blocking probability, average outage probability, packet loss probability and packet error rate (PER) for class j , respectively, with a certain admission control policy π .

The essence of the design problem is to derive an optimal connection admission control policy which is capable of maximizing the above system throughput, while simultaneously guaranteeing QoS requirements at physical, packet and connection levels.

In the following, first, we analyze the packet-level and physical-layer QoS requirements in terms of packet loss probability and outage probability, which are then passed to the connection level to decide the optimal connection admission control policy by formulating a constrained Markov decision process. In this sense, the connection admission control problem can be obtained by formulating a semi-Markov decision process (SMDP) problem.

4. Packet-level design

A system state \mathbf{s} is defined as $\mathbf{s} = [n_{a,1}, \dots, n_{a,j}]$, which represents the number of accepted users. In this section, we discuss how to choose the number of virtual channels $K_{s,j}$ for a given system state to guarantee the packet level QoS requirements in terms of packet loss probability. For simplicity, we first consider the case of no buffering, i.e., $B_j = 0$. The results are then extended to nonzero buffer sizes.

A. Departure rate with retransmissions

Without ARQ, the duration for a packet can be expressed as $\frac{N_p}{R_j}$, where N_p denotes the packet length in bits and R_j denotes the bit transmission rate. With ARQ, the packet duration, denoted by C_j , is the summation of the original packet duration and the duration for at most L_j retransmissions. As shown in [15], the mean duration can be expressed as

$$C_j = \frac{N_p}{R_j} (1 + (\rho_j)^{\frac{1}{L_j+1}} + \dots + (\rho_j)^{\frac{L_j}{L_j+1}}) \quad (3)$$

in seconds, where ρ_j denotes the target packet error rate for class j .

The packet departure rate for each virtual channel, denoted by $r_{d,j}$, can be obtained by

$$\begin{aligned} r_{d,j} &= \frac{1}{C_j} \\ &= \frac{\frac{R_j}{N_p}}{1 + (\rho_j)^{\frac{1}{L_j+1}} + \dots + (\rho_j)^{\frac{L_j}{L_j+1}}} \end{aligned} \quad (4)$$

in packets per second.

B. Packet loss probability

In the following, we assume that $B_j = 0$ and the incoming packets are allocated equally to the $K_{s,j}$ virtual channels, e.g., in a round-robin fashion. For each allocated virtual channel, the packet arrival rate can be expressed as $n_{o,j}r_{a,j}/K_{s,j}$, and the packet departure rate for each virtual channel, $r_{d,j}$, is given in (4).

To obtain the packet loss probability for given $n_{a,j}$, we first express the packet loss probability for a given $n_{o,j}$ as

$$P_L^j(n_{o,j}, K_{s,j}) = \begin{cases} 0 & \text{if } n_{o,j}r_{a,j} \leq K_{s,j}r_{d,j} \\ \frac{n_{o,j}r_{a,j} - K_{s,j}r_{d,j}}{n_{o,j}r_{a,j}} & \text{if } n_{o,j}r_{a,j} > K_{s,j}r_{d,j}. \end{cases} \quad (5)$$

Then the packet loss probability for a given $n_{a,j}$ can be obtained by

$$P_L^j(n_{a,j}, K_{s,j}) = \sum_{i=0}^{n_{a,j}} \text{Prob}\{n_{o,j} = i\} P_L^j(i, K_{s,j}) \quad (6)$$

$$\leq v_j \quad (7)$$

where v_j denotes the packet loss probability constraint, and $\text{Prob}\{n_{o,j} = i\}$ denotes the probability that i out of $n_{a,j}$ accepted users are in the ON state, which has Binomial distribution

$$\text{Prob}\{n_{o,j} = i\} = (p_{ON}^j)^i (1 - p_{ON}^j)^{n_{a,j}-i} \quad (8)$$

for $0 \leq i \leq n_{a,j}$.

C. Choosing $K_{s,j}$

In the above analysis, we assume that the packet generation traffic is modeled by an ON/OFF Markov process and buffer sizes are all zero. Under these assumptions, with a given number of accepted users $n_{a,j}$ and packet-level QoS constraints, $K_{s,j}$ is chosen to satisfy (7). For a general system, the virtual channel can be approximated by a $G/G/1/1 + B_j$ queue, where G denotes the generally distributed arrival and departure processes. Given a nonzero B_j , Equation (6) should be replaced by a corresponding packet loss probability formula by analyzing the $G/G/1/1+B_j$ queue, and then $K_{s,j}$ can be chosen to satisfy (7).

We note that for a given system state $\mathbf{s} = [n_{a,1}, \dots, n_{a,J}]$, an increase in the chosen $K_{s,j}$ can lead to improved packet-level performance. However, large $K_{s,j}$ introduces more mutual interference, which degrades the physical layer performance. The choice of $K_{s,j}$ represents a tradeoff between physical-layer and packet-level performances.

In the above, we only consider the packet-level QoS requirement in terms of packet loss probability. As discussed previously, other packet-level QoS requirements, such as packet access delay and delay jitter, can be satisfied by performing packet access control.

5. Physical-layer QoS: outage probability

Physical-layer performance is determined by the number of virtual channels, i.e., $K_{s,j}$. In the previous section, a lower bound of $K_{s,j}$ is given in (7), and an exact $K_{s,j}$ can then be determined by system resource allocation schemes, e.g., packet access control. In this section, we discuss how to ensure the physical-layer QoS requirements for beamforming systems in which $K_{s,j}$, where $j = 1, 2, \dots, J$, are known for each possible system state.

The QoS requirement in the physical layer can be represented by a target outage probability, defined as the probability that a target packet-error-rate (PER), or equivalently a target SINR,

cannot be satisfied. We consider two types of constraints: worst-state-outage-probability (WSOP) and average-outage-probability (AOP). The WSOP ensures that at any time instant and at any system state an outage probability constraint cannot be violated, while AOP only ensures a time-average outage probability constraint, which is less restrictive.

We first derive the outage probability for a given system state $\mathbf{s} = [n_{a,1}, \dots, n_{a,J}]$, in which a total of $\sum_{j=1}^J K_{s,j}$ channels are allocated. The outage probability for a given state is defined as the probability that a target PER, or equivalently a target SINR, cannot be satisfied. As shown in [17], the target SINR for a given PER constraint ρ_j , can be obtained as

$$\gamma_j = \frac{1}{g} [\ln a - \ln((\rho_j)^{\frac{1}{L_j+1}})] \tag{9}$$

in which a, g are constants depending on the chosen modulation and coding scheme [17]. Letting the SINR for an arbitrary packet k , where $k = 1, \dots, K$, given in (1) achieve its target value, we have the following matrix equation

$$[I_K - QF]\mathbf{p} = Q\mathbf{u} \tag{10}$$

where I_K is a K -dimensional identity matrix, power vector $\mathbf{p} = [p_1, \dots, p_K]^t$, $\mathbf{u} = \eta_0 B [1, \dots, 1]^t$, $(\cdot)^t$ denotes transpose, Q is a K -dimensional diagonal matrix with the i^{th} non-zero element as $\frac{\gamma_i R_i}{W}$, and F is a K by K matrix in which the element at the i^{th} row and the j^{th} column can

be expressed as $F_{ij} = \frac{\phi_{ij}^2}{\phi_{ii}^2}$.

To ensure a positive solution for power vector \mathbf{p} , we require the following feasibility condition,

$$\nu(QF) < 1 \tag{11}$$

where $\nu(\cdot)$ denotes the maximum eigenvalue, which is real-valued since the matrices are symmetric. Under the above feasibility condition, the power solution can be obtained by

$$\mathbf{p} = [I_K - QF]^{-1} Q\mathbf{u} \tag{12}$$

where $(\cdot)^{-1}$ denotes matrix inversion.

Therefore, the outage probability for a given system state \mathbf{s} in which $\sum_{j=1}^J K_{s,j}$ virtual channels are allocated, can be obtained as

$$\begin{aligned} P_{out}(\mathbf{s}) &= P_{out}(K_{s,1}, \dots, K_{s,J}) \\ &= \text{Prob}\{\nu(QF) \geq 1\} \end{aligned} \tag{13}$$

where $\text{Prob}\{A\}$ denotes the probability of event A .

Based on this state outage probability, the worst-state outage probability, denoted by P_{out}^{w} , and the average outage probability, denoted by P_{out}^{av} , can be expressed as follows

$$P_{out}^w = \max_{\mathbf{s} \in S} P_{out}(\mathbf{s}) \quad (14)$$

$$\leq \rho_w$$

$$P_{out}^{av} = \sum_{\mathbf{s} \in S} P_s P_{out}(\mathbf{s}) \quad (15)$$

$$\leq \rho_{av} \quad (16)$$

where ρ_w and ρ_{av} denote the WSOP and AOP constraints, respectively; P_s denotes the steady-state probability that the system is in state \mathbf{s} and S represents the set of all feasible system states, which will be discussed in Section VI.

6. Optimal connection admission control policy

The QoS requirements in the network layer can be characterized by blocking probability, defined as the probability that an incoming connection is blocked. The network-layer QoS as well as the other QoS should be guaranteed by a cross-layer connection admission control design.

In this chapter, we assume that the arrival process is Poisson distributed, the connection duration is exponentially distributed and the connection arrival and departure processes are independent. The system state is represented by the number of accepted connections. Under these assumptions, the process has the Markovian property that the future behavior of the process depends only on the present state and is independent of the past history [18]. In this sense, the connection admission control problem can be obtained by employing a SMDP approach.

A. SMDP components

A semi-Markov decision process includes the following components: system state, state space, action, action space, decision epoch, holding time, transition probability, policy and constraints. A brief description of the above SMDP components is summarized in Table I, and a detailed SMDP formulation can be found in [18].

System state is represented by the number of accepted connections, i.e., $\mathbf{s} = [n_{a,1}, \dots, n_{a,j}]$. A state is considered feasible if and only if this state can satisfy the worst-state-outage-probability and packet-loss-probability constraints. The state space includes all feasible system states, and can be expressed as

$$S = \{\mathbf{s}; P_{out}(\mathbf{s}) < \rho_w, \text{ and } P_L^j(n_{a,j}, K_{s,j}) \leq \nu_j, \text{ where } j = 1, \dots, J\}.$$

The formulation of the above state space can be summarized as follows:

- Compute the maximum number of accepted users for each class, denoted by M_j^{max} . The search procedure for M_j^{max} is presented in Figure 2;
- An enlarged state space, denoted by \bar{S} , can be defined as

$$\bar{S} = \left\{ \mathbf{s} = [n_{a,1}, \dots, n_{a,j}] : n_{a,j} \leq M_j^{max} \text{ for } j = 1, \dots, J \right\};$$

- The above \bar{S} can be truncated to the desired state space S as follows:
 - Initialize $S = \{\}$;
 - For each state $s \in \bar{S}$:
 - Choose appropriate $K_{s,j}$ for each j based on (7);
 - Evaluate $P_{out}(s)$ based on (13);
 - If $P_{out}(s) \leq \rho_w$, then $S = S + \{s\}$.
- We remark that in the above step, it is unnecessary to evaluate each system state in \bar{S} , since if $s \in S$, then all $s' \in \bar{S}$ such that $s' \leq s$ are also in S . Similarly, if s is not in S , then all $s' \in \bar{S}$ such that $s' \geq s$ are also not in S .

After formulating the state space, a virtual-channel-table can then be obtained via (7), which assigns the required number of virtual channels to each possible system state.

The state space, S , includes all the possible state vectors \mathbf{s} . The state space together with the SMDP constraints ensure the QoS requirements. Dynamic statistics can be characterized by expected holding time and transition probability. The expected holding time, denoted by $\tau_s(\mathbf{a})$, is the expected time until the next decision epoch after action \mathbf{a} is chosen in the present state \mathbf{s} . The transition probability, denoted by $p_{sy}(\mathbf{a})$, is the probability that the state at the next decision epoch is \mathbf{y} if action \mathbf{a} is selected at the current state \mathbf{s} .

For each given state $\mathbf{s} \in S$, an action $\mathbf{a} \in A_s$ is chosen according to a policy \mathbf{R} . A policy defines a mapping rule from the state space to the action space [7].

In the admission control problem discussed in this chapter, we have expressed QoS requirements in terms of blocking probability, packet loss probability, AOP and WSOP. While WSOP and packet loss probability requirements can be guaranteed by formulating the state space as shown in Table I, the other QoS requirements can be guaranteed by SMDP constraints.

B. Deriving an AC policy by linear programming

The policy can be chosen according to certain performance criterion, such as minimizing-blocking-probability or maximizing-throughput. Here we aim to find an optimal policy R^* which maximizes the throughput for any initial system state.

By formulating the admission problem as a SMDP, an optimal connection admission control policy can be obtained by using the decision variables z_{sa} , $\mathbf{s} \in S$, $\mathbf{a} \in A_s$, in solving the following linear programming (LP) problem [18]:

$$\max_{z_{sa} \geq 0, \mathbf{s}, \mathbf{a}} \sum_{\mathbf{s} \in S} \sum_{\mathbf{a} \in A_s} \sum_{j=1}^J \lambda_j a_j (1 - P_{out}(\mathbf{s})) P_{ON}^j r_{a,j} (1 - P_L^j) (1 - \rho_j) \tau_s(\mathbf{a}) z_{sa} \quad (17)$$

subject to the set of constraints

$$\begin{aligned} \sum_{\mathbf{a} \in A_m} z_{ma} - \sum_{\mathbf{s} \in S} \sum_{\mathbf{a} \in A_s} p_{sm}(\mathbf{a}) z_{sa} &= 0, m \in S \\ \sum_{\mathbf{s} \in S} \sum_{\mathbf{a} \in A_s} \tau_s(\mathbf{a}) z_{sa} &= 1 \\ \sum_{\mathbf{s} \in S} \sum_{\mathbf{a} \in A_s} (1 - a_j) \tau_s(\mathbf{a}) z_{sa} &\leq \Psi_j, j = 1, \dots, J \\ \sum_{\mathbf{s} \in S} \sum_{\mathbf{a} \in A_s} P_{out}(\mathbf{s}) \tau_s(\mathbf{a}) z_{sa} &\leq \rho_w \end{aligned}$$

where Ψ_j and ρ_{av} denotes the blocking probability and AOP constraints, respectively. In the above LP formulation, $\tau_s(\mathbf{a})z_{sa}$ represents the steady-state probability that the system is in state \mathbf{s} and an action \mathbf{a} is chosen. The objective function in (17) is to maximize the system throughput, the first constraint is the balance equation, and the second constraint ensures that the steady-state probabilities sum to one. The latter two constraints represent the QoS requirements in terms of blocking probability and average outage probability, respectively. Since the sample path constraints are included in the above linear programming approach, the optimal policy resulting from the SMDP is a randomized policy [7]: the optimal action $\mathbf{a}^* \in A_s$ for state \mathbf{s} , where A_s is the admissible action space, is chosen probabilistically according to the probabilities $z_{sa} / \sum_{\mathbf{a} \in A_s} z_{sa}$.

C. Implementation of the cross-layer connection admission control design

The cross-layer connection admission control design can be implemented as follows:

- Derive the connection admission control policy offline:
 - Formulate the state space according to the procedure in Section VI-A. Then derive a virtual channel table based on (7), which assigns a required number of virtual channels to each system state;
 - Formulate other SMDP components according to Table I;
 - The policy can then be derived according to (17);
 - Implement the connection admission control policy as a lookup table;
 - Whenever parameters change, repeat the above procedure to update the connection admission control lookup table and virtual channel table.
- Connection level implementation: whenever a connection arrives, the lookup table is employed to decide whether this packet can be accepted. The current state information, represented by the number of accepted users, and the virtual channel table, are then passed to the packet level.
- Packet level implementation:
 - The current state information and the virtual channel table are obtained from connection level;
 - For each system state, choose $K_{s,j}$, where $j = 1, \dots, J$, according to the virtual channel table;
 - For each incoming packet in class j , where $j = 1, \dots, J$, if the current number of simultaneously transmitted packets is less than $K_{s,j}$, the incoming packet can be transmitted. Otherwise, it is stored in the buffer;
 - The packets in the i^{th} virtual channel, where $i = 1, \dots, K_{s,j}$, are transmitted over the channel. An erroneous packet is retransmitted until it is correctly received or the maximum number of retransmissions is reached;
 - The chosen $K_{s,j}$, where $j = 1, \dots, J$, is passed to the physical layer.
- Physical layer implementation:
 - As discussed in packet-level implementation, $K_{s,j}$, where $j = 1, \dots, J$, is obtained from packet level;
 - Power is adjusted to the desired level, which is given in (12).

7. Numerical examples

In the following examples, we consider a packet-switched network with two classes of multimedia services. A circular antenna array and a uniformly distributed AoA are

assumed. QPSK and convolutionally coded modulation with rate $\frac{1}{2}$ and packet length $N_p = 1080$ is assumed at the transmitter. Under this scheme, the parameters of a and g in Equation (9) can be obtained from [17]. For simplicity, $B_1 = B_2 = 0$ is employed. Simulation parameters are summarized in Table II.

Without loss of generality, we choose $K_{s,j}$ to be the minimum number satisfying (7). The chosen $K_{s,j}$ can ensure the packet level QoS requirement while simultaneously minimizing the outage probability in the physical layer.

In the following, we first illustrate the performance for different packet loss probability constraints, in which the proposed policy and the policy for circuit-switched networks, discussed in [6], are compared. We then present the performance gain for the proposed connection admission control policy with ARQ over the system without ARQ schemes, such as the policies discussed in [8] [16].

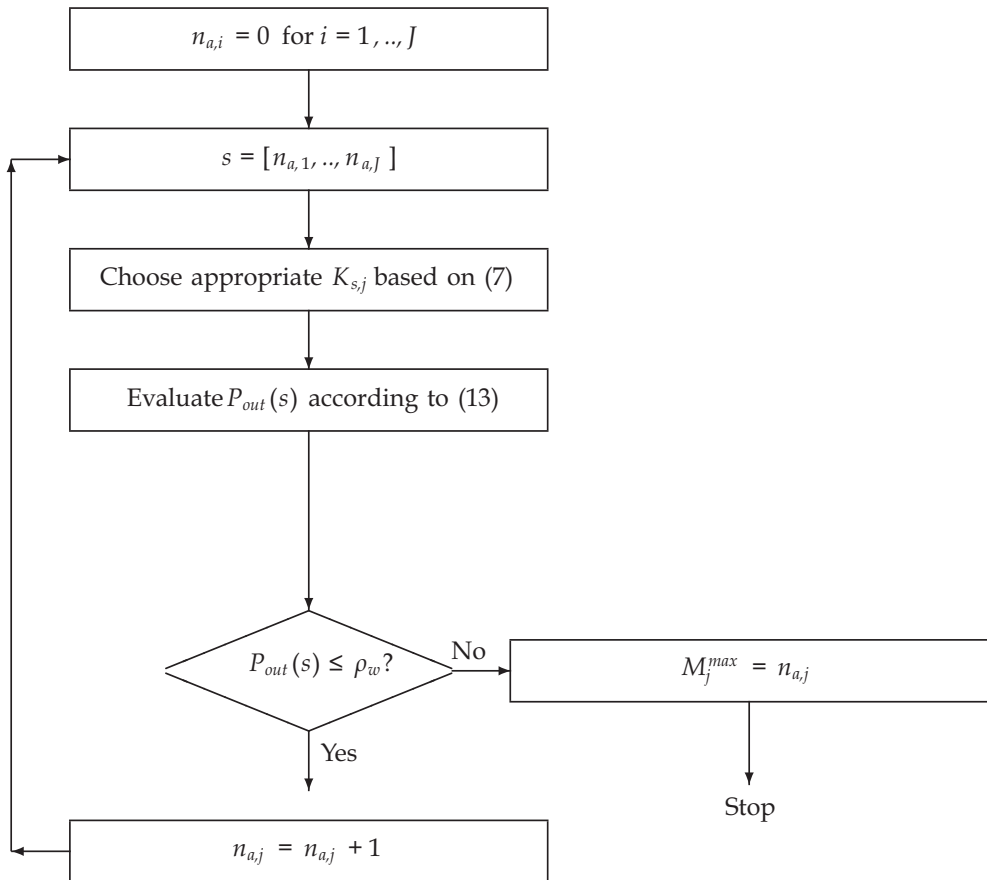


Fig. 2. Search procedure for M_j^{max} .

SMDP components	Notation	Expression
System state	\mathbf{s}	$\mathbf{s} = [n_{a,1}, \dots, n_{a,J}]$.
State space	S	$S = \{\mathbf{s}; P_{out}(K_{s,1}, \dots, K_{s,J}) < \rho_w,$ and $P_L^j(n_{a,j}, K_{s,j}) \leq \nu_j\}$.
Decision epochs	t_k	The set of all arrival and departure instances.
Action	\mathbf{a}	$\mathbf{a} = [a_1, \dots, a_J]$, where $a_j = 1$ represents the decision to accept a class j connection, while $a_j = 0$ represents a rejection.
Admissible action space	A_s	$A_s = \{\mathbf{a} : a_j = 0, \text{ if } \mathbf{s} + \mathbf{e}_s^j \notin S, \text{ and } \mathbf{a} \neq \mathbf{0} \text{ if } \mathbf{s} = \mathbf{0}\}$ in which \mathbf{e}_s^j represents a J -dimensional vector, which contains only zeros except for position j which contains a 1.
Expected holding time	$\tau_s(\mathbf{a})$	$\tau_s(\mathbf{a}) = \left(\sum_{j=1}^J \lambda_j a_j + \sum_{j=1}^J \mu_j n_s^j \right)^{-1}$.
Transition probability	$p_{sy}(\mathbf{a})$	$p_{sy}(\mathbf{a}) = \lambda_j a_j \tau_s(\mathbf{a})$, if $y = \mathbf{s} + \mathbf{e}_s^j$; and $p_{sy}(\mathbf{a}) = \mu_j n_s^j \tau_s(\mathbf{a})$, if $y = \mathbf{s} - \mathbf{e}_s^j$.
Policy	R	$R = \{R_s : S \rightarrow A R_s \in A_s, \forall \mathbf{s} \in S\}$ where A denotes the set of all admissible action space.
Constraints		$P_{out}^{av} \leq \rho_{av}$ and $P_b^j \leq \Psi_j$, where Ψ_j denotes the blocking probability constraint for class j .

Table 1. Formulating the optimal connection admission control problem as a SMDP.

W	3.84 MHz	a	90.2514
g	3.4998	γ_0	1.0942 dB
R_1	32 kbps	R_2	128 kbps
λ_1	0.01	λ_2	0.003
μ_1	0.005	μ_2	0.00125
$r_{a,1}$	50	$r_{a,2}$	200
P_{ON}^1	0.4	P_{ON}^2	0.6
ρ_w	0.5	M	2

Table 2. Simulation parameters.

A. Performance of a packet-switched network

In the following, we compare the performance for different packet loss probability constraints, in which no ARQ schemes are employed. Since a strict packet loss probability constraint introduces a large blocking probability, which may lead to infeasibility in (18), we now relax the blocking probability constraints to 0.5 for both classes to ensure problem feasibility. The target SINR for class 1 and class 2 users are set to 10 and 7 dB, respectively.

Figures 3-6 compare the blocking probability, average outage probability, average packet loss probability and system throughput for different packet-loss-probability constraints, respectively. For simplicity, we assume the packet loss probability constraints are the same for both classes, which are denoted by P_{loss} constraint in the figures. From these figures, we observe that the performance in one layer strongly depends on the QoS constraints of the other layers. For example, given an average outage probability constraint, relaxing the packet-loss-probability constraint can dramatically reduce the blocking probability in the

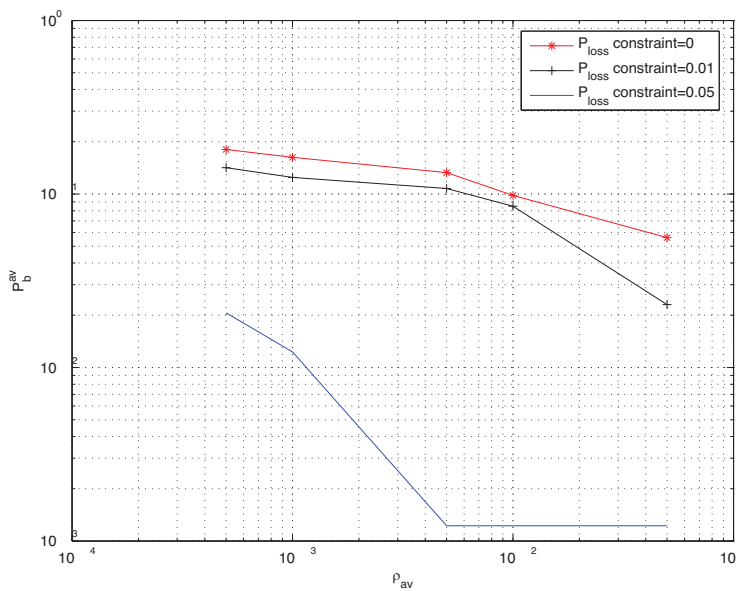


Fig. 3. Blocking probability as a function of ρ_{av} .

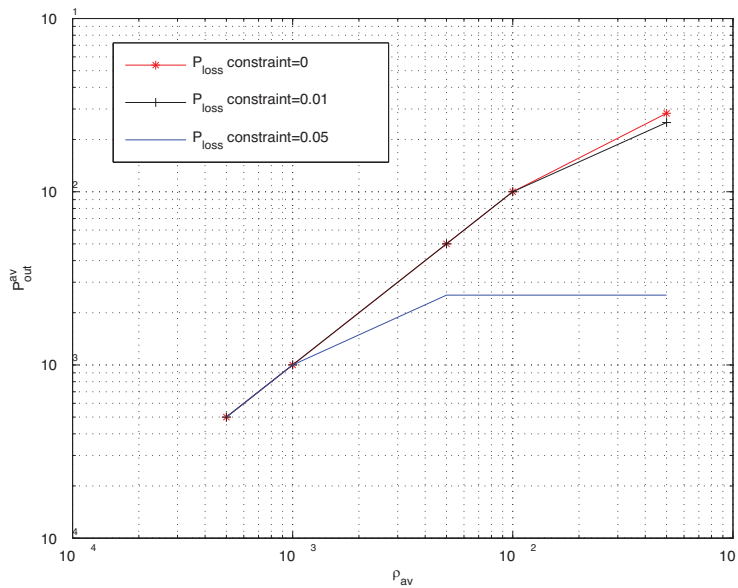


Fig. 4. Outage probability as a function of ρ_{av} .

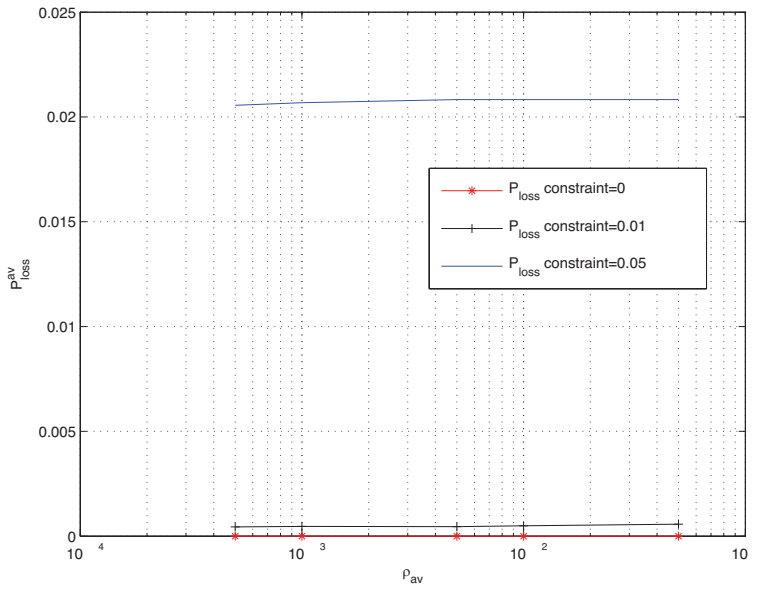


Fig. 5. Average packet loss probability as a function of ρ_{av} .

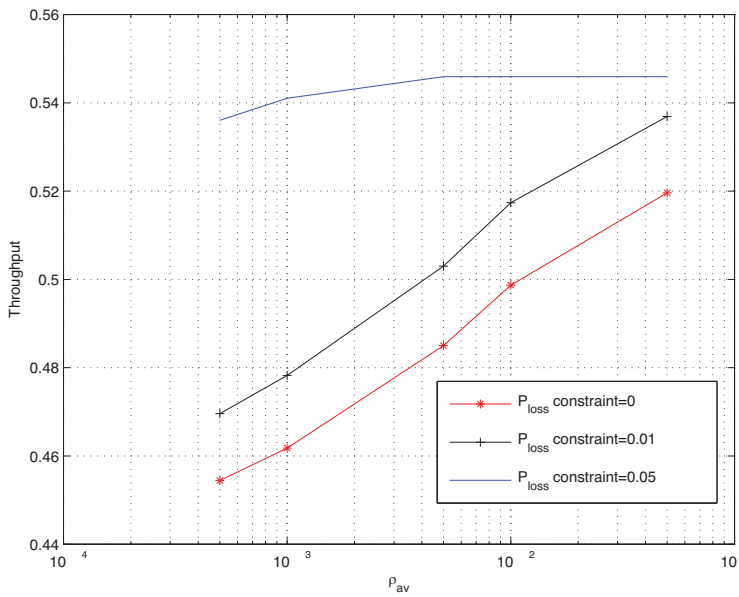


Fig. 6. Throughput as a function of ρ_{av} .

network layer, while simultaneously improving the overall system throughput. This can be explained by the fact that with a given physical layer performance, a large packet loss probability constraint allows more users to access the network. In the system we investigate, with $\rho_{av} = 10^{-2}$, relaxing the packet loss probability constraint from 0 to 0.05 can reduce the blocking probability from 10^{-1} to 10^{-3} , i.e., by 99%, while improving the throughput from 0.5 to 0.545, i.e., by 9%.

We note that the achieved packet loss probability in Figure 5 is obtained by averaging the measurements over a long-term period, while $P_{loss\ constraint}$ denotes the maximum allowed packet loss probability for each system state.

With a CAC policy in a circuit-switched network, e.g., the work discussed in [6], a zero packet-loss-probability can be ensured. As observed in Figures 3-6, in a packetized system which allows a non-zero packet loss probability, this zero packet loss probability leads to an inefficient utilization of the system resource and as a result degrades the connection level performance as well as the overall system throughput.

B. Performance by employing packet retransmissions

Figures 7-9 compare the performance between a system without ARQ, e.g., [8] [16], and a system with ARQ. In these figures, ARQ = i is equivalent to $L_1 = L_2 = i$. The blocking probability is set to 0.1 for both classes and the target overall PERs are set to $\rho_1 = 10^{-4}$ and $\rho_2 = 10^{-6}$, respectively. The packet loss probability constraints are set to 0.05 for both classes.

From Figure 7, it is observed that with ARQ, the blocking probability and outage probability can be reduced. This represents a tradeoff between transmission delay and system performance. For example, with $\rho_{av} = 10^{-3}$, employing an ARQ scheme with $L_j = 1$ can decrease the blocking probability from 10^{-3} to 10^{-4} , i.e., by 90%, while simultaneously reducing the outage probability from 10^{-3} to almost 10^{-6} , i.e., by 99%.

In the above, we have studied the physical and network layer performance by employing ARQ. We now investigate how ARQ schemes affect the packet level performance. As shown in (4), with an increased L_j , the departure rate is decreased due to retransmissions, which increases the packet loss probability. However, at the same time, an increased L_j also reduces the transmission error, allowing more virtual channels simultaneously presented in the system, which in turn decreases the packet loss probability. Therefore, the packet loss probability is determined by the above positive and negative impacts of ARQ. If the positive impact dominates, the packet loss probability is reduced by employing ARQ, as shown in the upper figure in Figure 8. Otherwise, if the negative impact dominates, the packet loss probability is degraded by employing ARQ, as shown in the lower figure in Figure 8. We note that the above degradation is not very significant. As shown in Figure 9, by employing ARQ, the overall system throughput can be improved.

Although increasing L_j may further improve system performance, it dramatically increases the computational complexity of the SMDP-based connection admission control policy. In [15], it has been shown that when L_j exceeds a certain level, further increasing L_j cannot improve the performance significantly. Therefore, there is no need to choose a large L_j . A detailed discussion on the impact of ARQ and how to choose L_j can be found in [15], in which a packet-level AC is discussed which employs an ARQ-based algorithm to reduce probability of outage. In this chapter, we have only addressed the connection admission control policy for a given L_j . The optimization of L_j is beyond the scope of this discussion.

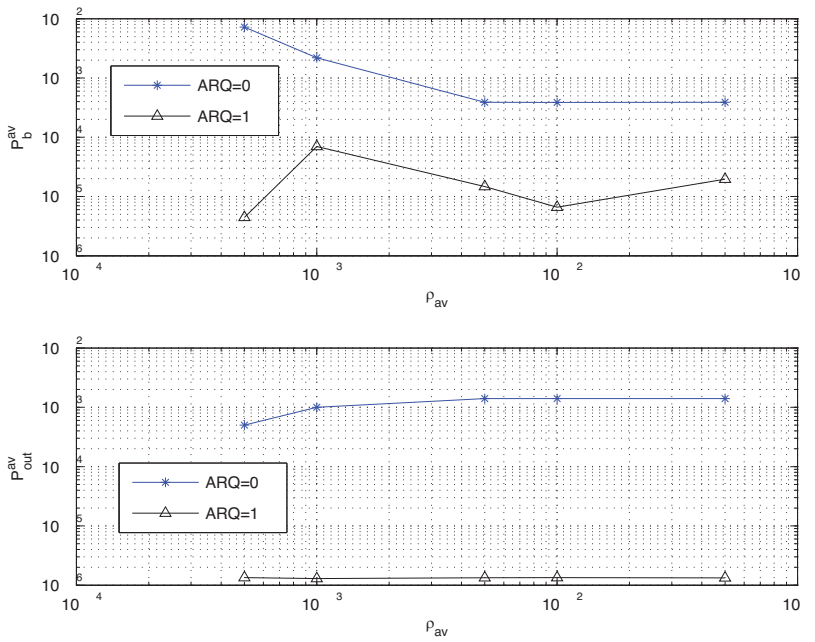


Fig. 7. Blocking and outage probabilities as a function of ρ_{av} .

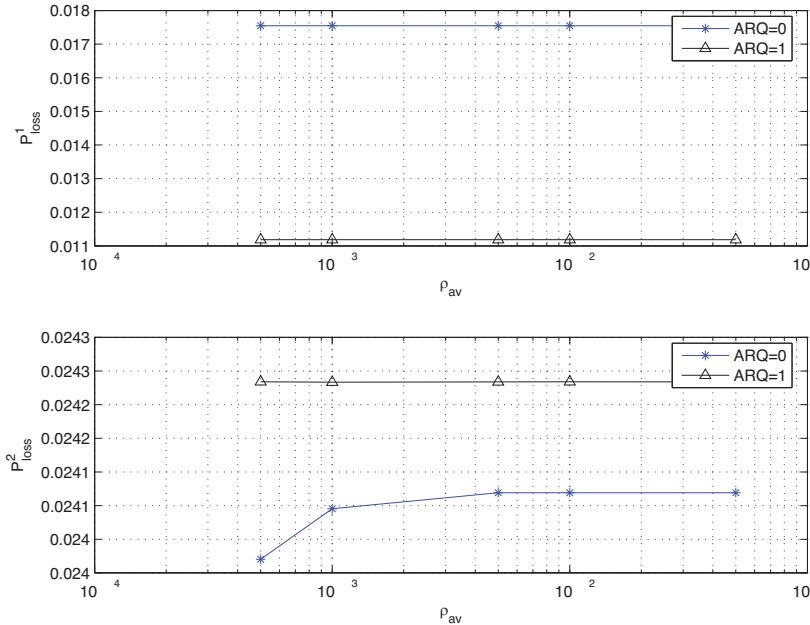


Fig. 8. Packet loss probability as a function of ρ_{av} .

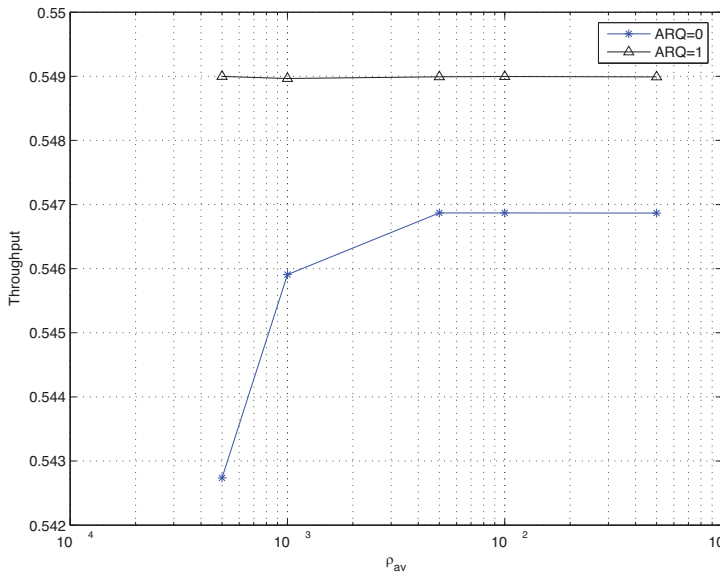


Fig. 9. Throughput as a function of ρ_{av} .

8. Summary

In summary, this chapter provides a framework for joint optimization of packet-switched multiple-antenna systems across physical, packet and connection levels. We extend the existing CAC policies in packet-switched networks to more general cases, where the SINR may vary quickly relative to the connection time, as encountered in multiple antenna base stations. Compared with the CAC policy for circuit-switched networks, the proposed connection admission control policy allows dynamical allocation of limited resources, and as a result, is capable of efficient resource utilization. The proposed CAC policy demonstrates a flexible method of handling heterogeneous QoS requirements while simultaneously optimizing overall system performance.

9. References

- [1] R. M. Rao, C. Comaniciu, T.V. Lakshman and H. V. Poor, "Call admission control in wireless multimedia networks", *IEEE Signal Processing Magazine*, pp. 51-58, September 2004.
- [2] Y. Kwok and V. K. N. Lau, "On admission control and scheduling of multimedia burst data for CDMA systems", *Wireless Networks*, pp. 495-506, 2002, Kluwer Academic Publishers.
- [3] S. Brueck, E. Jugl, H. Ketschau, M. Link, J. Mueckenheim, and A. Zaporozhets, "Radio Resource Management in HSDPA and HSUPA", *Bell Labs Technical Journal*, 11(4), pp. 151-167, 2007.

- [4] S. Singh, V. Krishnamurthy, and H. V. Poor, "Integrated voice/Data call admission control for wireless DS-CDMA systems", *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1483-1495, June 2002.
- [5] C. Comaniciu and H. V. Poor, "Jointly optimal power and admission control for delay sensitive traffic in CDMA networks with LMMSE receivers", *IEEE Trans. Signal Processing*, vol. 51, no. 8, pp. 2031-2042, August 2003.
- [6] W. Sheng and S. D. Blostein, "A Maximum-Throughput Call Admission Control Policy for CDMA Beamforming Systems", *Proc. IEEE WCNC 2008*, Las Vegas, March 2008, pp. 2986-2991.
- [7] F. Yu, V. Krishnamurthy, and V. C. M Leung, "Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless CDMA networks", *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 542-555, February 2006.
- [8] K. Li and X. Wang, "Cross-layer optimization for LDPC-coded multirate multiuser systems with QoS constraints", *IEEE Trans. Signal Processing*, vol. 54, no. 7, pp. 2567-2578, July 2006.
- [9] I. E. Telatar, "Capacity of multi-antenna Gaussian channels", Technical Report, AT&T Bell Labs, 1995.
- [10] S. D. Blostein and H. Leib, "Multiple antenna systems: Role and impact in future wireless access", *Communication Magazine*, vol. 41, no. 7, pp. 94-101, July 2003.
- [11] Y. Hara, "Call admission control algorithm for CDMA systems with adaptive antennas", *IEEE Proc. Veh. Technol. Conf.*, pp. 2518-2522, May 2000.
- [12] K. I. Pedersen and P. E. Mogensen, "Directional power-based admission control for WCDMA systems using beamforming antenna array systems", *IEEE Trans. Vehicular Technology*, vol. 51, no. 6, pp. 1294-1303, November 2002.
- [13] F. R. Farrokhi, L. Tassiulas and K. J. R. Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays", *IEEE Trans. Communications*, vol. 46, no. 10, pp. 1313-1324, October 1998.
- [14] A. M. Wyglinski and S. D. Blostein, "On uplink CDMA cell capacity: mutual coupling and scattering effects on beamforming", *IEEE Trans. Vehicular Technology*, vol. 52, no. 2, pp. 289-304, March 2003.
- [15] W. Sheng and S. D. Blostein, "Cross-layer Admission Control Policy for CDMA Beamforming Systems", *EURASIP Journal on Wireless Communications and Networking, Special Issue on Smart Antennas*, July 2007.
- [16] L. Wang and W. Zhuang, "A call admission control scheme for packet data in CDMA cellular communications", *IEEE Trans. Wireless Communications*, vol. 5, no. 2, pp. 406-416, February 2006.
- [17] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links", *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, September 2004.
- [18] H. C. Tijms, *Stochastic Modelling and Analysis: a Computational Approach*, U.K.: Wiley, 1986.

Advanced Access Schemes for Future Broadband Wireless Networks

Gueguen Cédric and Baey Sébastien
*Université Pierre et Marie Curie (UPMC) - Paris 6
France*

1. Introduction

Bandwidth allocation in next generation broadband wireless networks (4G systems) is a difficult issue. The scheduling shall support efficient multimedia transmission services which require managing users mobility with fairness while increasing system capacity together. The past decades have witnessed intense research efforts on wireless communications. In contrast with wired communications, wireless transmissions are subject to many channel impairments such as path loss, shadowing and multipath fading. These phenomena severely affect the transmission capabilities and in turn the QoS experienced by applications, in terms of data integrity but also in terms of the supplementary delays or packet losses which appear when the effective bit rate at the physical layer is low.

Among all candidate transmission techniques for broadband transmission, Orthogonal Frequency Division Multiplexing (OFDM) has emerged as the most promising physical layer technique for its capacity to efficiently reduce the harmful effects of multipath fading. This technique is already widely implemented in most recent wireless systems like 802.11a/g or 802.16. The basic principle of OFDM for fighting the effects of multipath propagation is to subdivide the available channel bandwidth in sub-frequency bands of width inferior to the coherence bandwidth of the channel (inverse of the delay spread). The transmission of a high speed signal on a broadband frequency selective channel is then substituted with the transmission on multiple subcarriers of slow speed signals which are very resistant to intersymbol interference and subject to flat fading. This subdivision of the overall bandwidth in multiple channels provides frequency diversity which added to time and multiuser diversity may result in a very spectrally efficient system subject to an adequate scheduling.

The MAC protocols currently used in wireless local area networks were originally and primarily designed in the wired local area network context. These conventional access methods like Round Robin (RR) and Random Access (RA) are not well adapted to the wireless environment and provide poor throughput. More recently intense research efforts have been given in order to propose efficient schedulers for OFDM based networks and especially opportunistic schedulers which preferably allocate the resources to the active mobile(s) with the most favourable channel conditions at a given time. These schedulers take benefit of multiuser and frequency diversity in order to maximize the system throughput. In fact, they highly rely on diversity for offering their good performances. The higher the diversity the more efficient are these schedulers, the less the multiuser diversity

the more underachieved they are. However, in this context, the challenge is to avoid fairness deficiencies owing to unequal spatial positioning of the mobiles in order to guarantee QoS whatever the motion of the mobile in the cell. Indeed, since the farther mobiles have a lower spectral efficiency than the closer ones due to pathloss, the mobiles do not all benefit of an equal priority and average throughput which induces unequal delays and QoS (Fig. 1).

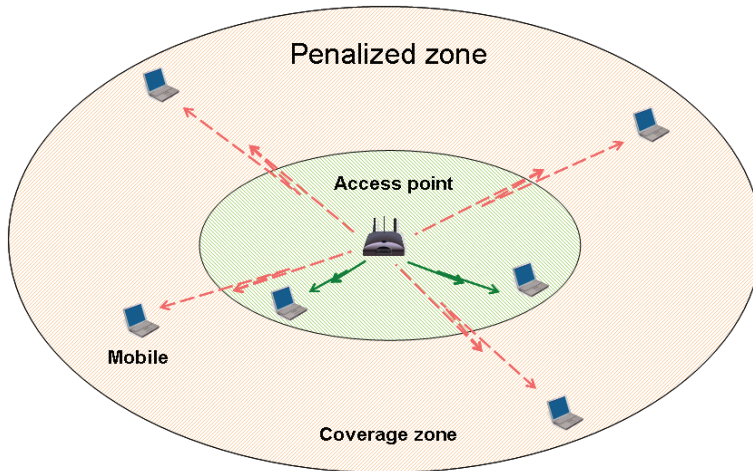


Fig. 1. Illustration of opportunistic scheduling fairness issue.

2. Multiuser OFDM system description

In this chapter, we focus on the proper allocation of radio resources among the set of mobiles situated in the coverage zone of an access point both in the uplink and in the downlink. The scheduling is performed in a centralized approach. The packets originating from the backhaul network are buffered in the access point which schedules the downlink transmissions. In the uplink, the mobiles signal their traffic backlog to the access point which builds the uplink resource mapping.

The physical layer is operated using an OFDM frame structure compliant to the OFDM mode of the IEEE 802.16-2004 (Hoymann, 2005). The total available bandwidth is divided in sub-frequency bands or subcarriers. The radio resource is further divided in the time domain in frames. Each frame is itself divided in time slots of constant duration. The time slot duration is an integer multiple of the OFDM symbol duration. The number of subcarriers is chosen so that the width of each sub-frequency band is inferior to the coherence bandwidth of the channel. Moreover, the frame duration is fixed to a value much smaller than the coherence time (inverse of the Doppler spread) of the channel. With these assumptions, the transmission on each subcarrier is subject to flat fading with a channel state that can be considered static during each frame.

The elementary resource unit (RU) is defined as any (subcarrier, time slot) pair. Each of these RUs may be allocated to any mobile with a specific modulation order. Transmissions performed on different RUs by different mobiles have independent channel state variations (Andrews et al., 2001). On each RU, the modulation scheme is QAM with a modulation order adapted to the channel state between the access point and the mobile to which it is

allocated. This provides the flexible resource allocation framework required for opportunistic scheduling.

The frame structure supposed a perfect time and frequency synchronization between the mobiles and the access point as described in (Van de Beek et al., 1999). Additionally, perfect knowledge of the channel state is supposed to be available at the receiver (Li et al., 1999). The current channel attenuation on each subcarrier and for each mobile is estimated by the access node based on the SNR of the signal sent by each mobile during the uplink contention subframe. Assuming that the channel state is stable on a scale of 50 ms (Truman & Brodersen, 1997), and using a frame duration of 2 ms, the mobiles shall transmit their control information alternatively on each subcarrier so that the access node may refresh the channel state information once every 25 frames

3. Scheduling techniques in OFDM wireless networks

This chapter focuses on the two major scheduling techniques which have emerged in the literature: Maximum Signal-to-Noise Ratio (MaxSNR), Proportional Fair (PF). Furthermore, it will present an improvement of PF scheduling which avoid fairness deficiencies: the Compensated Proportional Fair (CPF).

3.1 Classical scheduling: Round Robin

Before studying opportunistic schedulers, we bring to mind the characteristics of classical schedulers. Round Robin (RR) (Nagle, 1987; Kuurne & Miettinen, 2004) is a well-attested bandwidth allocation strategy in wireless networks. RR allocates an equal share of the bandwidth to each mobile in a ring fashion. However, it does not take in consideration that far mobiles have a much lower spectral efficiency than closer ones which does not provide full fairness. Moreover, the RR does not take benefit of multiuser diversity which results in a bad utilization of the bandwidth and in turn, poor system throughput.

3.2 Maximum Signal-to-Noise Ratio

Many schemes are derived from the Maximum Signal-to-Noise Ratio (MaxSNR) technique (also known as Maximum Carrier to Interference ratio (MaxC/I)) (Knopp & Humblet, 1995; Wong et al., 1999; Wang & Xiang, 2006). MaxSNR exploits the concept of opportunistic scheduling. Priority is given to the mobile which currently has the greatest signal-to-noise ratio (SNR). Profiting of the multiuser diversity and continuously allocating the radio resource to the mobile with the best spectral efficiency, MaxSNR strongly improves the system throughput. It dynamically adapts the modulation and coding to allow always making the most efficient use of the radio resource and coming closer to the Shannon limit. However, a negative side effect of this strategy is that the closest mobiles to the access point have disproportionate priorities over mobiles more distant since their path loss attenuation is much smaller. This results in a severe lack of fairness as illustrated in Fig. 1.

3.3 Proportional Fair

Proportional Fair (PF) algorithms have recently been proposed to incorporate a certain level of fairness while keeping the benefits of multiuser diversity (Viswanath et al., 2002; Kim et al., 2002; Anchun et al., 2003; Svedman et al., 2004; Kim et al., 2004). In PF based schemes, the basic principle is to allocate the bandwidth resources to a mobile when its channel

conditions are the most favourable with respect to its time average. At a short time scale, path loss variations are negligible and channel state variations are mainly due to multipath fading, statistically similar for all mobiles. Thus, PF provides an equal sharing of the total available bandwidth among the mobiles as RR. Applying the opportunistic scheduling technique, system throughput maximization is also obtained as with MaxSNR. PF actually combines the advantages of the classical schemes and currently appears as the best bandwidth management scheme.

In PF-based schemes, fairness consists in guaranteeing an equal share of the total available bandwidth to each mobile, whatever its position or channel conditions. However, since the farther mobiles have a lower spectral efficiency than the closer ones due to pathloss, all mobiles do not all benefit of an equal average throughput despite they all obtain an equal share of bandwidth. This induces heterogeneous delays and unequal QoS. (Choi & Bahk, 2007; Gueguen & Baey, 2009; Holtzman, 2001) demonstrate that fairness issues persist in PF-based protocols when mobiles have unequal spatial positioning.

3.4 Compensated Proportional Fair

This QoS and fairness issues can be solved by an improvement of the PF called Compensated Proportional Fair (CPF). CPF introduces correction factors in the PF in order to compensate the path loss negative effect on fairness while keeping the PF system throughput maximization properties. With this compensation, CPF is aware of the path loss disastrous effect on fairness and adequate priorities between the mobiles are always adjusted in order to ensure them an equal throughput. This scheduling finely and simultaneously manages all mobiles. Keeping a maximum number of flows active across time but with relatively low traffic backlogs, CPF is designed for best profiting of the multiuser diversity taking advantage of the dynamics of the multiplexed traffics. Thus, preserving the multiuser diversity, CPF takes a maximal benefit of the opportunistic scheduling technique and maximizes the system capacity better than MaxSNR and PF access schemes. Well-combining the system capacity maximization and fairness objectives required for 4G OFDM wireless networks, an efficient support of multimedia services is provided.

At each scheduling epoch, the scheduler computes the maximum number of bits $B_{k,n}$ that can be transmitted in a time slot of subcarrier n if assigned to mobile k , for all k and all n . This number of bits is limited by two main factors: the data integrity requirement and the supported modulation orders.

The bit error probability is upper bounded by the symbol error probability (Wong & Cheng, 1999) and the time slot duration is assumed equal to the duration T_s of an OFDM symbol. The required received power $P_r(q)$ for transmitting q bits in a resource unit while keeping below the data integrity requirement BER_{target} is a function of the modulation type, its order and the single-sided power spectral density of noise N_0 . For QAM and a modulation order M on a flat fading channel (Proakis, 1995):

$$P_r(q) = \frac{2N_0}{3T_s} \left[\operatorname{erfc}^{-1} \left(\frac{BER_{target}}{2} \right) \right]^2 (M-1), \quad (1)$$

where $M = 2^q$ and erfc is the complementary error function. $P_r(q)$ may also be determined in practice based on BER history and updated according to information collected on experienced BER. Additionally, the transmit power $P_{k,n}$ of mobile k on subcarrier n is upper bounded to a value P_{max} which complies with the transmit Power Spectral Density regulation:

$$P_{k,n} \leq P_{\max}. \quad (2)$$

Given the channel gain $a_{k,n}$ experienced by mobile k on subcarrier n (including path loss and multipath fading):

$$P_r(q) \leq a_{k,n} P_{\max}. \quad (3)$$

The channel gain model on each subcarrier considers free space path loss a_k and multipath Rayleigh fading $\alpha_{k,n}^2$ (Parsons, 1992):

$$a_{k,n} = a_k \alpha_{k,n}^2. \quad (4)$$

a_k is dependent on the distance between the access point and mobile k . $\alpha_{k,n}^2$ represents the flat fading experienced by mobile k on subcarrier n . $\alpha_{k,n}$ is Rayleigh distributed with an expectancy equal to unity. Consequently, the maximum number of bits $q_{k,n}$ of mobile k which can be transmitted on a time slot of subcarrier n while keeping below its BER target is:

$$q_{k,n} \leq \left\lceil \log_2 \left(1 + \frac{3P_{\max} T_s a_k \alpha_{k,n}^2}{2N_0 \left[\operatorname{erfc}^{-1} \left(\frac{\operatorname{BER}_{\text{target}}}{2} \right) \right]^2} \right) \right\rceil. \quad (5)$$

We further assume that the supported QAM modulation orders are limited such as q belongs to the set $S = \{0, 2, 4, \dots, q_{\max}\}$. Hence, the maximum number of bits $B_{k,n}$ that will be transmitted on a time slot of subcarrier n if this resource unit is allocated to the mobile k is:

$$B_{k,n} = \max \{ q \in S, q \leq q_{k,n} \}. \quad (6)$$

At each scheduling epoch and for each time slot, MaxSNR based schemes allocate the subcarrier n to the active mobile k which has the greatest $B_{k,n}$ value while the PF scheme consists in allocating the subcarrier n to the mobile k which has the greatest factor $F_{k,n}$ defined as:

$$F_{k,n} = \frac{B_{k,n}}{R_{k,n}}, \quad (7)$$

where $R_{k,n}$ is the time average of the $B_{k,n}$ values. However, considering rounded $B_{k,n}$ values in the allocation process have a negative discretization side effect on the PF performances. Several mobiles may actually have a same $F_{k,n}$ value with significantly different channel state with respect to their time average. More accuracy is needed in the subcarrier allocation process for prioritizing the mobiles. It is more profitable to allocate the subcarrier n to the mobile k which has the greatest $f_{k,n}$ value defined by:

$$f_{k,n} = \frac{b_{k,n}}{r_{k,n}}, \quad (8)$$

where:

$$b_{k,n} \leq \log_2 \left(1 + \frac{3P_{\max} T_s a_k \alpha_{k,n}^2}{2N_0 \left[\operatorname{erfc}^{-1} \left(\frac{\operatorname{BER}_{\text{target}}}{2} \right) \right]^2} \right), \quad (9)$$

and $r_{k,n}$ is the $b_{k,n}$ average over a sliding time window.

PF outperforms MaxSNR providing an equal system capacity and partially improving the fairness (Gueguen & Baey, 2009). Based on the PF scheme, this chapter presents a new scheduler that achieves high fairness while preserving the system throughput maximization. It introduces a parameter called "Compensation Factor" (CF_k), that takes into account the current path loss impact on the average achievable bit rate of the mobile k . It is defined by:

$$CF_k = \frac{b_{\text{ref}}}{b_k}. \quad (10)$$

b_{ref} is a reference number of bits that may be transmitted on a subcarrier considering a reference free space path loss a_{ref} for a reference distance d_{ref} to the access point and a multipath fading equal to unity:

$$b_{\text{ref}} \leq \log_2 \left(1 + \frac{3P_{\max} T_s a_{\text{ref}}}{2N_0 \left[\operatorname{erfc}^{-1} \left(\frac{\operatorname{BER}_{\text{target}}}{2} \right) \right]^2} \right). \quad (11)$$

b_k represents the same quantity but considering a distance d_k to the access point:

$$b_k \leq \log_2 \left(1 + \frac{3P_{\max} T_s a_{\text{ref}} \left(\frac{d_{\text{ref}}}{d_k} \right)^\beta}{2N_0 \left[\operatorname{erfc}^{-1} \left(\frac{\operatorname{BER}_{\text{target}}}{2} \right) \right]^2} \right), \quad (12)$$

with β the experienced path loss exponent.

The distance d_k of the mobile k to the access point is evaluated thanks to the channel state estimation time average (Jones & Raleigh, 1998). The CPF scheduling consists then in allocating a time slot of subcarrier n to the mobile k which has the greatest $CPF_{k,n}$ value:

$$CPF_{k,n} = f_{k,n} CF_k = \left(\frac{b_{k,n}}{r_{k,n}} \right) CF_k. \quad (13)$$

The CPF scheduling algorithm is detailed in Fig. 2. The distance correction factor CF_k adequately compensates the lower spectral efficiencies of far mobiles and the resulting

$CPF_{k,n}$ parameters bring high fairness in the allocation process. Far mobiles get access to the resource more often than close mobiles and inverse proportionally to their spectral efficiency. Thereby, an equal throughput is provided to each mobile. Moreover, CPF also keeps the PF opportunistic scheduling advantages thanks to the $f_{k,n}$ parameters which take into account the channel state. In contrast with MaxSNR and PF which satisfy much faster the mobiles which are close to the access point, the CPF keeps more mobiles active but with a relatively low traffic backlog. Satisfaction of delay constraints is more uniform and, preserving the multiuser diversity, a better usage of the bandwidth is made. This jointly ensures fairness and system throughput maximization.

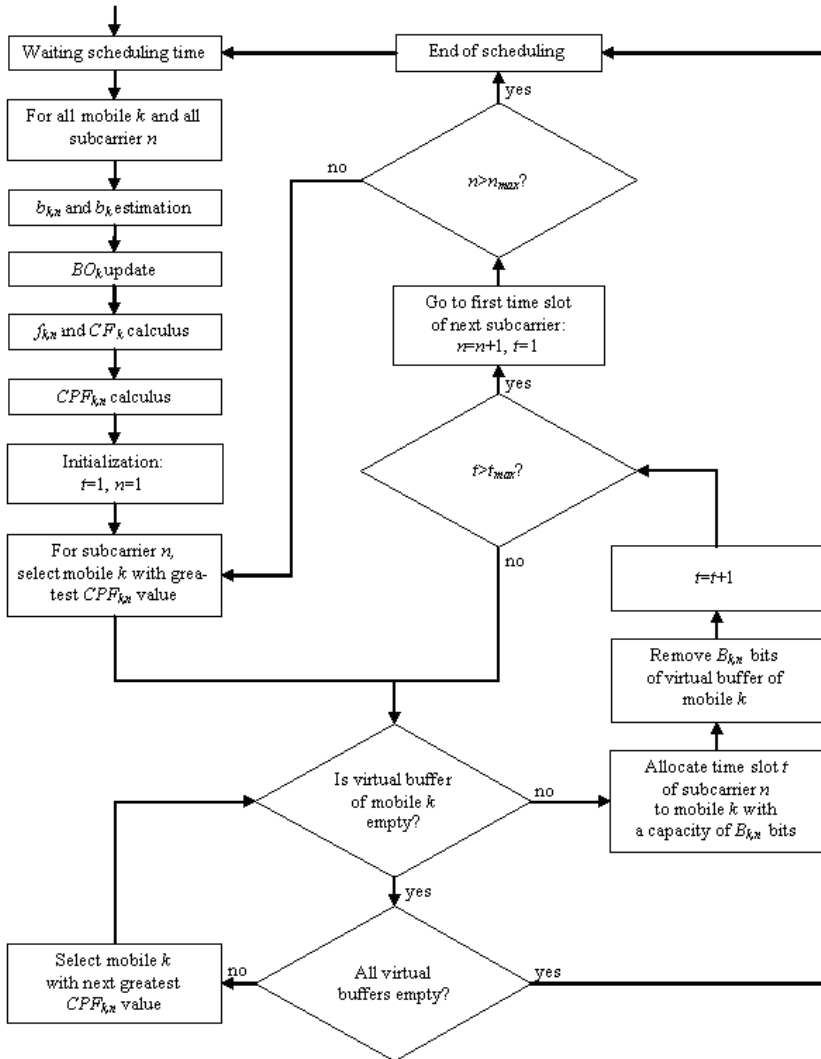


Fig. 2. CPF scheduling algorithm flow chart.

4. Performance evaluation

In this section an extend performance evaluation using OPNET discrete event simulations is proposed. We focus on two essential performance criteria: fairness and offered system capacity.

In the simulations, a frame is composed of 5 time slots and 128 subcarriers. β is assumed equal to 2 and the maximum transmit power satisfies:

$$10\log_{10}\left(\frac{P_{\max}T_s}{N_0} \times a_{ref}\right) = 24 \text{ dB} . \quad (14)$$

All mobiles run a videoconference application. The traffic is composed of an MPEG-4 video stream (Baey, 2004) multiplexed with an AMR voice stream (Brady, 1969). This demanding type of application generates a high volume of data with high sporadicity and requires tight delay constraints which substantially complicates the task of the scheduler. The average bit rate of each source is 80 Kbps. The traffic load is set by varying the number of mobiles. This allows to study the ability of each scheduler to take advantage of the multiuser diversity.

A crucial objective for modern multiple access schemes is the full support of multimedia transmission services. Evaluating the QoS offered by a scheduling scheme should not only focus on the classical delay and jitter analysis. Indeed, a meaningful constraint regarding delay is the limitation of the occurrences of large values. In this aim, we define the concept of *delay outage* by analogy with the concept of outage used in system coverage planning. A mobile transmission is in delay outage when its packets experience a delay greater than a given threshold. The delay experienced by each mobile is tracked all along the lifetime of its connection. At each transmission of a packet of mobile k , the ratio of the total number of packets whose delay exceeded the threshold divided by the total number of packets transmitted since the beginning of the connection is computed. The result is called Packet Delay Outage Ratio (PDOR) of mobile k and is denoted $PDOR_k$. Fig. 3 illustrates an example cumulative distribution of the packet delay of a mobile at a given time instant.

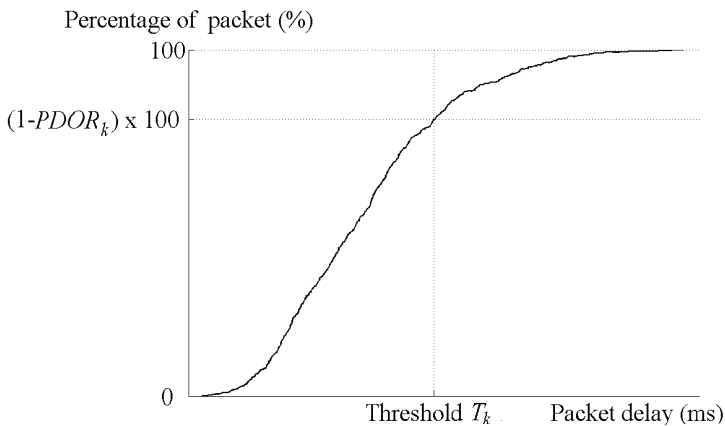


Fig. 3. Example packet delay CDF and experienced PDOR.

The PDOR target is defined as the maximum ratio of packets of mobile k that may be delivered after its delay threshold T_k . This characterizes the delay requirements of any mobile in a generic approach. In the following, the PDOR target is set to 5 % and the threshold time T_k is fixed to the value of 80 ms considering real time constraints. The BER_{target} value is taken equal to 10^{-3} .

Note that the problem we are studying in this chapter is quite different with the sum-rate maximization with water-filling for instance. The purpose of the schedulers presented in this chapter is to maximize the traffic load that can be admitted in the wireless access network while fulfilling delay constraints. This is achieved by both taking into account the radio conditions but also the variations in the incoming traffic. In this context, it cannot be assumed for instance that each mobile has some traffic to send at each scheduling epoch. Traffic overload is not realistic in a wireless access network because it corresponds to situations where the excess traffic experiences an unbounded delay. This is why, in the showed simulations, the traffic load (offered traffic) does not exceed the system capacity. In these conditions the offered traffic is strictly equal to the traffic carried over the wireless interface and all mobiles get served sooner or later. The bit rate sent by each mobile is equal to its incoming traffic. Fairness in terms of bit rate sent by each mobile is rigorously achieved. The purpose of the scheduler is to dynamically assign the resource units to the mobiles at the best time in order to meet the traffic delay constraints. This is why the PDOR is adopted as a measure of the fairness in terms of QoS level obtained by each mobile.

4.1 Static scenario

In order to study the influence of the distance on the scheduling performances, a first half of mobiles are positioned close to the access point at a distance of $1.5 d_{ref}$. The second half of mobiles are twice over farther. With these settings, the values of $B_{k,n}$ for the two groups of mobiles are respectively 4 and 2 bits when $\alpha_{k,n}^2$ equals unity.

Fairness is the most difficult objective to reach. It consists in ensuring the same ratio of packets in delay outage to all mobiles, below the PDOR target. Fig. 4 displays the overall PDOR for various traffic loads. The influence of distance on the scheduling is also studied.

Classical RR yields bad results (Fig. 4a). Indeed, since multiuser diversity is not exploited, the overall spectral efficiency is small and system throughput is low. Consequently, the delay targets are exceeded as soon as the traffic load increases. Based on opportunistic scheduling, MaxSNR (Fig. 4b), PF (Fig. 4c) and CPF (Fig. 4d) provide better system performances. However, with MaxSNR and PF, close mobiles easily respect their delay requirement but the farther experience much higher delays and go beyond their PDOR target when the traffic load increases. This shows their difficulty to ensure fairness when the mobiles have heterogeneous positions. Indeed, with MaxSNR, unnecessary priorities are given to close mobiles who easily respect their QoS constraints while more attention should be given to the farther. These inadequate priority management dramatically increases the global mobile PDOR and mobile dissatisfaction. PF brings slightly more fairness and allocates more priority to far mobiles. The result on global overall PDOR indicates that some flows can be slightly delayed to the benefit of others without significantly affecting their QoS.

The CPF was built on this idea. The easy satisfaction of close mobiles (with better spectral efficiency) offers a degree of freedom which ideally should be exploited in order to help the farther ones. CPF dynamically adapts the priorities function of the mobile location. This results in allocating to each mobile the accurate share of bandwidth required for the

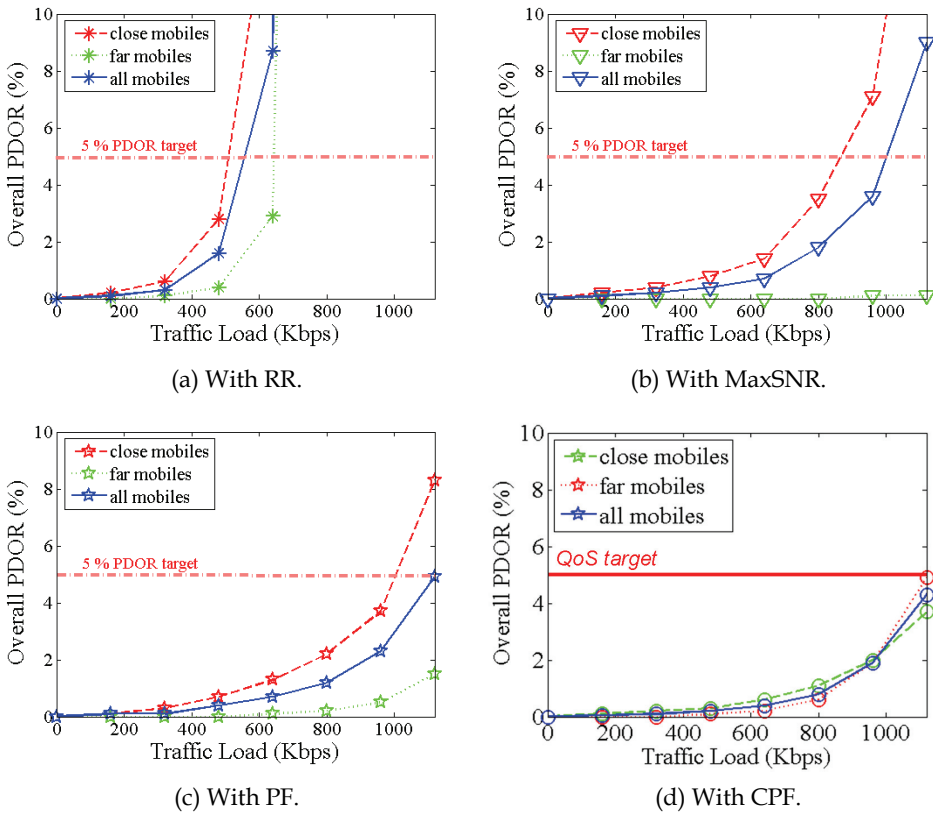


Fig. 4. Measured QoS with respect to distance.

satisfaction of its QoS constraints, whatever its position. Like this, the problem of fairness is solved with CPF which provides comparable QoS levels to all mobiles whatever their respective location and allows to reach higher traffic loads with an acceptable PDOR (below the PDOR target). Additionally, observing the global PDOR value (for all mobiles), we can notice that, besides ensuring high fairness, CPF provides a better overall QoS level as well.

Fig. 5 shows the average number of bits carried per allocated Resource Unit by each tested scheduler under various traffic loads. Looking at the cost of this high fairness and mobile satisfaction in terms of system capacity, it appears that no system throughput reduction has been done with CPF. As expected, the non opportunistic Round Robin scheduling provides a constant spectral efficiency, i.e. an equal bit rate per subcarrier whatever the traffic load since it does not take advantage of the multiuser diversity. The three other tested schedulers show better results. In contrast with RR, with the opportunistic schedulers (MaxSNR, PF, CPF), we observe a characteristic inflection of the spectral efficiency curves when the traffic load increases. Exploiting the supplementary multiuser diversity, the system capacity is highly extended. This result also shows that the CPF scheduling has slightly better performances than the two other opportunistic schedulers. This improved multiplexing efficiency is obtained by processing all service flows jointly and opportunistically. Keeping

more mobiles active but with a relatively lower traffic backlog, the CPF scheme preserves multiuser diversity and takes more advantage of it obtaining a slightly higher bit rate per subcarrier (cf. Fig. 5).

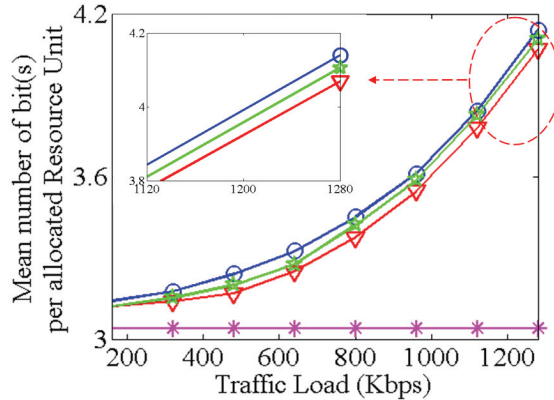


Fig. 5. Bandwidth usage efficiency.

The performance of the four schedulers can be further qualified by computing the theoretical maximal system throughput. Considering the Rayleigh distribution, it can be noticed that $\alpha_{k,n}^2$ is greater or equal to 8 with a probability of only 0.002. In these ideal situations, close mobiles can transmit/receive 6 bits per RU while far mobiles may transmit/receive 4 bits per RU. If the scheduler always allocated the RUs to the mobiles in these ideal situations, an overall efficiency of 5 bits per RU would be obtained which yields a theoretical maximal system throughput of 1600 Kbps. Comparing this value to the highest supported traffic load of 1280 Kbps (cf. Fig. 5) further demonstrates the good efficiency obtained with the opportunistic schedulers that nearly always serve the mobiles when their channel conditions are very good with near to 4.2 bits per allocated subcarrier.

4.2 Mobile scenario

In the above scenario, the mobiles are static, and positioned at two distinct locations. The objective was to demonstrate the opportunistic behaviour of the schedulers and also clearly exhibit their ability to provide fairness whatever the respective position of the mobiles. This second scenario brings additional results in a more general context that includes mobility. We constituted two groups of 7 mobiles that both move straight across the cell, following the pattern described in Fig. 6 and Fig.7. Each mobile has a speed of 3 km/h and the cell radius is taken equal to 5 km ($3 d_{ref}$). When a group of mobiles comes closer to the access point, the other group simultaneously goes farther away.

Considering the path loss, the Rayleigh fading and this mobility model, we have computed in Fig. 8 the evolution of the mean number of bits that may be transmitted per Resource Unit for each group of mobiles, averaging over all the Resource Units of a frame. This shows the impact of the mobile position on the mean $m_{k,n}$ values. Fig. 9 reports the mean number of bit(s) per “allocated” Resource Unit for each group of mobiles (RR performances are not presented here since RR is not able to support such a high traffic load.). The results underline the ability of opportunistic schedulers to take advantage of the multiuser diversity in order to maximize the

spectral efficiency. With opportunistic scheduling, a Resource Unit is allocated only when the associated channel state is good and the number of bits that may be transmitted is greater than the mean. This provides high system throughput with a mean number of bits per allocated Resource Unit varying between 3 and 5 (Fig. 9) while the average Resource Unit capacity ranges between only 1.8 and 3.5 (Fig. 8). This also further confirms the results of Fig.5: CPF offers slightly better results than MaxSNR and PF in terms of spectral efficiency.

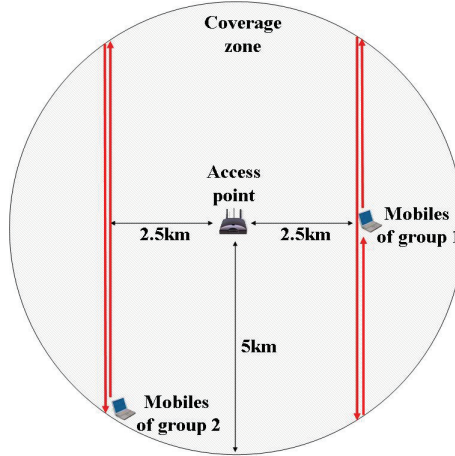


Fig. 6. Mobility pattern.

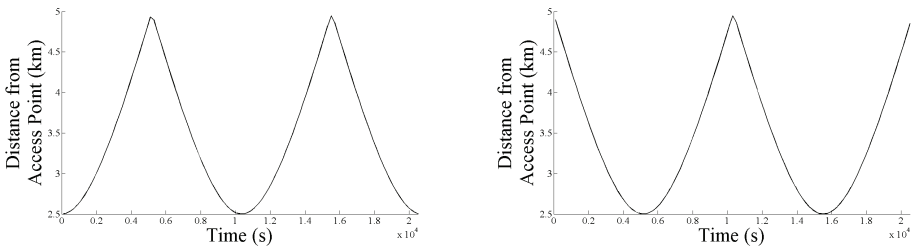


Fig. 7. Position of the mobiles across time (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

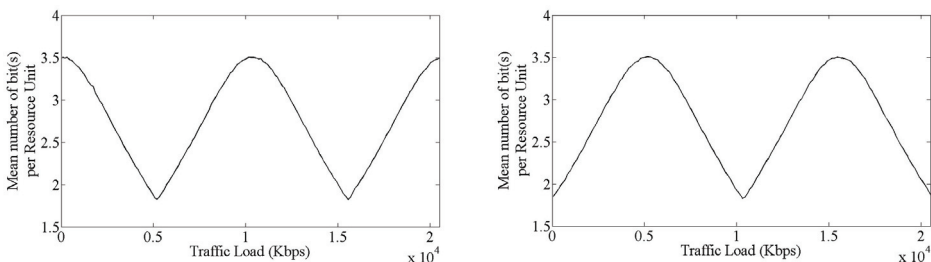


Fig. 8. Mean number of bit(s) per Resource Unit for each group of mobiles (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

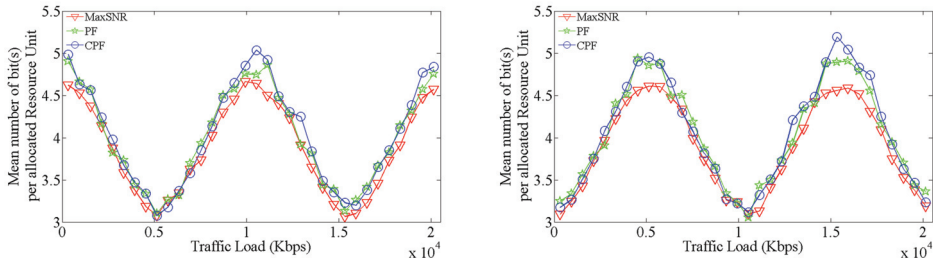


Fig. 9. Mean number of bit(s) per allocated Resource Unit for each group of mobiles (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

Regarding fairness, Fig. 10 reports the mean delay experienced by the transmitted packets of each group of mobiles across time. MaxSNR is highly unfair. Indeed, as soon as the mobiles move away from the access point, they experience a very high delay. PF offers better results. It brings more fairness and globally attenuates the delay peaks. However, we observe that

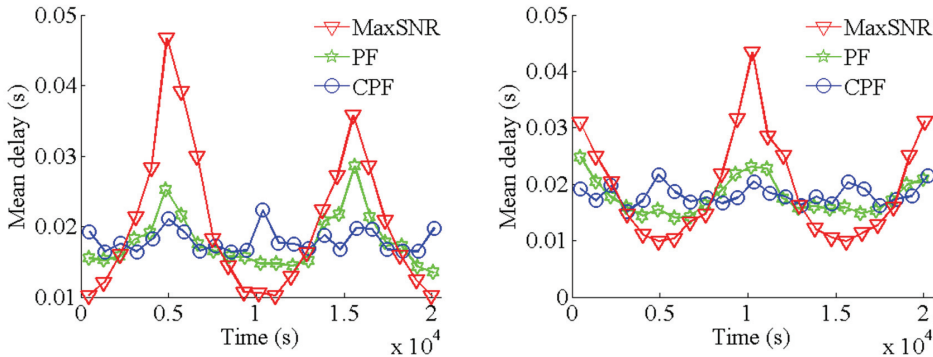


Fig. 10. Mean delay experienced by each group of mobiles (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

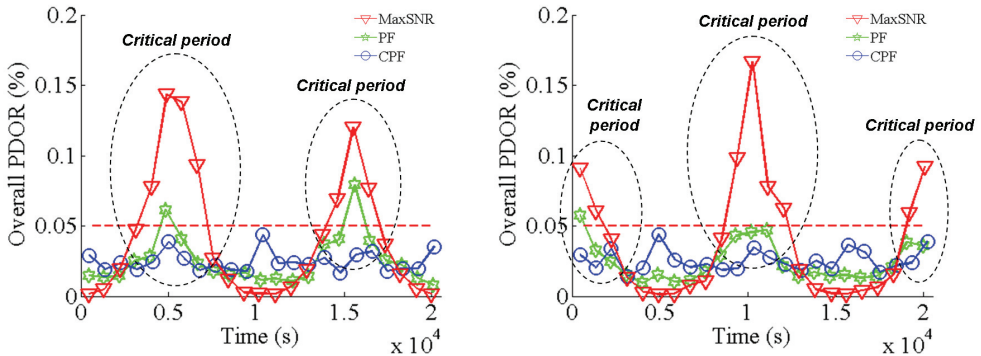


Fig. 11. PDOR fluctuations experienced by each group of mobiles (for mobiles of group 1 on the left and for mobiles of group 2 on the right).

CPF is the one that most smoothes the delay peaks. CPF continuously allocates the adequate priorities between the mobiles considering their relative movement across the cell. Providing a totally fair allocation of the bandwidth resources, CPF smoothes the delay experienced by each mobile across time.

Fig. 11 shows the mean PDOR experienced by each group of mobiles across the time. As we can see in Fig. 10 and Fig. 11, there is a very high correlation between the mean packet delay and the mean ratio of packets delivered after the delay threshold (mean PDOR). Reducing the magnitude of the delay peaks, PF greatly improves the mobile satisfaction with a greater reactivity than MaxSNR in critical periods. CPF further enhances the PF performances. It dynamically adjusts the priority of the mobile considering its position so that the PDOR values are further decreased. This results in a very fair resource allocation that fully satisfies the delay constraints whatever the motion of the mobile.

5. Conclusion

In the literature, several scheduling schemes have been proposed for maximizing the system throughput. However, guaranteeing a high fairness appeared as unfeasible without sacrificing system capacity. In this chapter, we have presented an improvement of the PF scheduling scheme yet acknowledged as the most promising so far. This scheme, called "Compensated Proportional Fair (CPF)", allows to avoid the tradeoff between fairness and system capacity. It has a low complexity and is easily implementable on all OFDM based networks like 802.11a/g and 802.16 networks. CPF sparingly delays the flows of close mobiles with good spectral efficiency in order to favor the flows of the farther mobiles which need more attention for fulfilling their delay constraints. Performance results show that CPF provides both high fairness and system throughput maximization making a better usage of multiuser diversity.

6. References

- Hoymann, C. (2005). Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16. *Computer Networks*, Vol. 49, No. 3, pp. 341-363, ISSN: 1389-1286
- Andrews, M.; Kumaran, K.; Ramanan, K. Stolvar, A. & whiting P. (2001). Providing quality of service over a shared wireless link. *IEEE Communications Magazine*, Vol. 39, No.2, pp. 150-154, ISSN: 0163-6804
- Van de Beek, J.-J. ; Borjesson, P.O. ; Boucheret, M.-J ; Landstrom, D. ; Arenas, J.-M. Odling, P.; Ostberg, C.; Wahlgvist, M. & Wilson, S.K. (1999). A time and frequency synchronization scheme for multiuser OFDM. *IEEE J.Sel. Areas Commun*, Vol.17, No.11, pp 1900-1914, ISSN: 0733-8716
- Li, Y.G; Seshadri, N. & Ariyavisitakul, S. (1999). Channel estimation for ofdm systems with transmitter diversity in mobile wireless channels. *IEEE J.Sel. Areas Commun*, Vol.17, No.3, pp 461-471, ISSN: 0733-8716
- Truman, T.E. & Brodersen, R.W (1997). A measurement-based characterization of the time variation of an indoor wireless channel, *proceedings of Int. Universal Personal Communications Record (ICUPC)*, pp. 25-32, ISBN: 0-7803-3777-8, San Diego, CA, USA

- Nagle (1987), On packet switches with infinite storage, *IEEE Transactions on Communications*, vol. 35, no. 4, pp. 435 - 438
- Kuurne and Miettinen (2004), Weighted round robin scheduling strategies in (E)GPRS radio interface, in *Proc. IEEE Int. Vehicular Technology Conference (VTC)*, vol. 5, pp. 3155 - 3159
- Knopp, R. & Humblet, P. (1995). Choi, J. (1996). Information capacity and power control in single-cell multiuser communications, *proceedings of IEEE Conference on Communications (ICC)*, pp 331-335, ISBN: 0-7803-2486-2, Seattle, WA, USA
- Wong, C.Y.; Cheng, R.S.; Lataief, K.B. & Murch, R.D. (1999). Multiuser OFDM with adaptative subcarrier, bit and power allocation. *IEEE J.Sel. Areas Commun*, Vol.17, No.10, pp 1747-1758, ISSN: 0733-8716
- Wang, X. & Xiang, Y. (2006). An OFDM-TDMA/SA MAC protocol with QoS constaints for broadband wireless LANs. *ACM/Springer Wireless Networks*, Vol. 12, No. 2, pp 159-170, ISSN: 1022-0038
- Viswanath, P.; Tse, D.N.C. & Laroia, R. (2002). Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, Vol. 48, No.6, pp. 1277-1294, ISSN: 0018-9448
- Kim, H.; Kim, K.; Han, Y. & Lee, J. (2002). An efficient scheduling algorithm for QoS in wireless packet data transmission, *proceedings of IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp 2244-2248, ISBN: 0-7803-7589-0, Lisboa, Portugal
- Anchun, W.; Liang, X.; Xjiin, S.X. & Yan, Y. (2003). Dynamic resource management in the fourth generation wireless systems, *proceedings of IEEE Int. Conference on Communication Technology (ICCT)*, pp 1095-1098, ISBN: 7-5635-0686-1, Beijing, China
- Svedman, P.; Wilson, K. & Ottersen, B. (2004). A QoS-aware proportional fair scheduler for oportunistic OFDM, *proceedings of IEEE Int. Vehicular Technology Conference (VTC)*, pp 558-562, ISBN: 0-7803-8521-7, Los angeles, CA, USA
- Kim, H.; Kim, K.; Han, Y. & Yun, S. (2004). A proportional fair scheduling for multicarrier transmission systems, *proceedings of IEEE Int. Vehicular Technology Conference (VTC)*, pp. 409-413, ISBN: 0-7803-8521-7, Los angeles, CA, USA
- Choi, J.-G. & Bahk, S. (2007). Cell-throughput analysis of the proportional fair scheduler in the single-cell environment. *IEEE Transactions on Vehicular Technology*, vol. 56, No.2, pp. 766-778, ISSN: 0018-9545
- Gueguen, C. & Baey, S. (2009). A Fair Opportunistic Access Scheme for Multiuser OFDM Wireless Networks. *EURASIP Journal on Wireless Communications and Networking. Special issue on "Fairness in Radio Resource Management for Wireless Networks"*. Article ID 726495, pp. 70-83
- Holtzman, J. (2001). Asymptotic analysis of proportional fair algorithm, *proceedings of IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 33-37, ISBN: 0-7803-7244-1, San Diego, CA, USA
- Proakis, J.G. (1995), *Digital Communications*. 3rd ed. New York: McGraw-Hill
- Parsons, J.D (1992). *The Mobile Radio Propagation Channel*, Wiley, ISBN: 978-0-471-98857-1
- Jones and Raleigh (1998), Channel estimation for wireless ofdm systems, *Proceedings of IEEE Int. Global Telecommunications Conference (GLOBECOM)*, vol. 2, pp. 980-985
- Baey, S. (2004). Modeling MPEG4 video traffic based on a customization of the DBMAP, *proceedings of Int. Symposium on Performance Evaluation of Computer and*

Telecommunication Systems (SPECTS), pp. 705-714, ISBN: 1-56555-284-9, San Jose, California, USA

Brady, P. (1969). A model for generating on-off speech patterns in two-conversation. *Bell System Technical Journal*, vol. 48, No.1, pp. 2445-2472

Medium Access Control in Distributed Wireless Networks

Jun Peng

*University of Texas - Pan American, Edinburg, Texas
United States of America*

1. Introduction

Medium access control (MAC) is a fundamental and challenging problem in networking. This problem is at the data link layer which interfaces the physical layer and the upper layers. A solution to this problem in a particular network thus needs to factor in the characteristics of the physical layer and the upper layers, which makes the MAC problem both a challenging and evolving problem. Medium access control in distributed wireless networks is one of the most active research areas in networking because distributed wireless networks are diverse and evolving fast.

One of the most well-known problem in medium access control in distributed wireless networks is the hidden terminal problem. Hidden terminals are interesting but problematic phenomena in distributed wireless networks. Basically, even if two nodes in a wireless network cannot sense each other, they may still cause collisions at the receiver of each other (1). If the hidden terminal problem is not well addressed, a wireless network may have a significantly degraded performance in every aspect, since frequent packet collisions consume all types of network resources such as energy, bandwidth, and computing power but generate no useful output.

There are basically two existing approaches to the hidden terminal problem. One is the use of an out-of-band control channel for signaling a busy data channel when a packet is in the air (2; 3; 4; 5). This approach is effective in dealing with hidden terminals but requires an additional control channel. The more popular approach to the hidden terminal problem is the use of in-band control frames for reserving the medium before a packet is transmitted (6; 7; 8; 9; 10). The popular IEEE 802.11 standard (11) uses this approach in its distributed coordination function (DCF).

Basically, before an IEEE 802.11 node in the DCF mode transmits a packet to another node, it first sends out a Request to Send (RTS) frame after proper backoffs and deferrals. If the receiver successfully receives the RTS frame and the channel is clear, the receiver responds with a Clear to Send (CTS) frame, which includes a Duration field informing its neighbors to back off during the specified period. In an ideal case, the hidden terminals of the initiating sender will successfully receive the CTS frame and thus not initiate new transmissions when the packet is being transmitted.

However, control frames have limited effectiveness in dealing with hidden terminals because they may not be able to reach all the intended receivers due to signal attenuation, fading, or interference (12). In addition, control frames have considerably long airtimes

because they are recommended to be transmitted at the basic link rate in both narrow-band and broadband IEEE 802.11 systems. Moreover, they have relatively long physical layer preambles and headers. In-band control frames therefore introduce significant network overhead, even though they do not use an out-of-band control channel.

This article introduces a new approach of *bit-free* control frames to addressing the disadvantages of the traditional control frames. Basically, with the new approach, control information is carried by the *airtimes* instead of the *bits* of control frames. The airtime of a frame is robust against interference and channel effects. In addition, a bit-free control frame carries no meaningful bits so that no preamble or header is needed for it (Section 6 presents a fundamental view on bit-free control frames).

In investigating the performance of the new approach, we have first analyzed the potential performance gains of the IEEE 802.11 DCF if its traditional control frames are replaced by bit-free control frames. We have then modified the original protocol with the new approach of bit-free control frames and done extensive simulations. Our investigation has shown that the modified protocol improves the average throughput of a wireless network from fifteen percent to more than one hundred percent.

The rest of this article is organized as follows. Section 2 introduces our observations and analysis. Section 3 presents our modifications to the IEEE 802.11 DCF. We then show in Section 4 the comprehensive simulation results comparing the modified protocol to the original one. We introduce the related work in Section 5 and a fundamental view on the presented approach in Section 6. Finally, we give our conclusions in Section 7.

2. Observations and analysis

Our first observation is that the CTS frame of an IEEE 802.11 node may not be able to reach all the hidden terminals of the initiating sender, which was also studied in some related work such as (12). One source of the problem is that recovering the bits in a frame is a delicate process so that to corrupt a frame being received by a node is usually much easier than to correctly receive a frame from the same node. In general, if a node is receiving a frame at the power level L , then another node may corrupt the frame by generating a power of level l at the receiver in the channel that is several times lower than L . In particular, when $\frac{l}{L}$ is lower than the "capture" power ratio threshold, then the frame will be corrupted.

An example is shown in Fig. 1. We assume in the example that the network is a homogeneous network, which means that all the nodes are the same in terms of parameters such as transmission power and receive/carrier sense power thresholds. We also assume that the signal power deteriorates at a rate of $(\frac{1}{d})^4$ where d is the propagation distance (i.e., the receiver is beyond the crossover distance from the sender), the carrier sense range of a node is twice of its transmission range r , and the capture power ratio threshold is 10, as used as the default settings in ns-2 and in some other studies (12). Under these assumptions, node C shown in Fig. 1 is a hidden terminal to node A. Meanwhile, node C cannot correctly receive a frame from node B, since it is out of node B's transmission range. However, node C can still corrupt a frame at node B that is from node A. Therefore, node C is a hidden terminal of node A that cannot be addressed by the CTS control frame sent by node B. Actually, all nodes falling into the closed region enclosing node C are hidden terminals of node A that cannot be addressed by the CTS frames of node B.

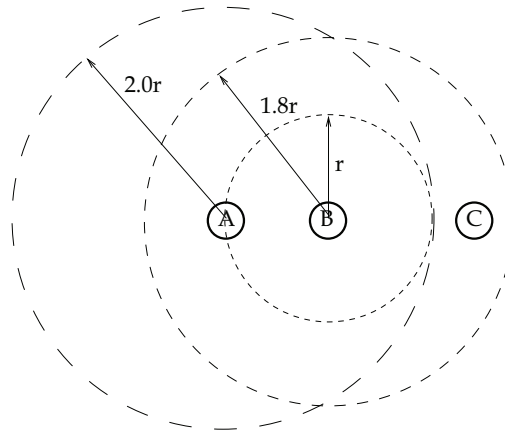


Fig. 1. A Case of a Failed CTS Frame for Reserving the Medium

Besides their limited effectiveness in dealing with hidden terminals, the control frames of IEEE 802.11 DCF also introduce significant overhead. There are two factors increasing the overhead. First, the control frames are recommended in both narrow-band and broadband IEEE 802.11 systems to be transmitted at the basic link rate for rate compatibility among competing nodes, which makes the bits in a control frame “flow” relatively slowly. Second, a bit-based frame, whatever the number of payload bits in it, needs a physical layer preamble and header for successful bit delivery.

As specified in IEEE 802.11, a DSSS (Direct Sequence Spread Spectrum) physical layer introduces 192-bit overhead (144-bit preamble plus 48-bit header) to each frame, while a FHSS (Frequency-Hopping Spread Spectrum) physical layer has an overhead of 128 bits (96-bit preamble plus 32-bit header). In the DSSS case, a RTS frame only uses 36% of its air time for delivering specific MAC information. It is even worse for a CTS frame, for which the percentage is 26%. The situation is relatively better in the FHSS case. The percentages are, however, still low at 44% and 33% for a RTS frame and a CTS frame, respectively.

We may use some analysis to demonstrate how a protocol that overcomes the two disadvantages of the IEEE 802.11 DCF may decrease the control overhead and thus improve the throughput of a network. After proper deferrals and backoffs, an IEEE 802.11 sender in the DCF mode starts to transmit the RTS frame. With a probability of p_c , however, the RTS frame may encounter a contention collision because another contending sender may have drawn a similar backoff delay. Even if there is no contention collision, the RTS frame may still face a collision later because of the possible existence of hidden terminals. We may assume the probability of such a collision as p_h . Therefore, a RTS frame with a transmission time of t_{rts} consumes a medium time of

$$T_{rts} = \frac{1}{(1 - p_c) \times (1 - p_h)} \times (T_{bo} + t_{rts}) \quad (1)$$

before it is successfully received by the intended receiver, where T_{bo} is the average backoff time in a contention and the interframe space times are considered as negligible.

If the RTS frame is successfully received by the intended receiver, we may assume that the CTS frame will not have a collision at the initiating sender, considering that the RTS frame

has already reserved the medium around the initiating sender. However, there is still a probability of $f \times p_h$ (f is the hidden terminal residual factor of DCF and $f \leq 1$) that the data packet may encounter a collision because some hidden terminals of the initiating sender may have failed to receive the CTS frame, as explained earlier. When the data packet has a collision, the RTS/CTS/Data process needs to be repeated. If we denote the transmission time of a CTS frame and of an ACK frame by t_{cts} and t_{ack} , respectively, then the medium time consumed for delivering a data packet and all its retransmissions is as follows

$$T = \frac{1}{1 - f \times p_h} \times (T_{rts} + t_{cts} + t_{data}) + t_{ack}. \quad (2)$$

We also assume here that the ACK may not have a collision, as in the CTS frame case.

The average time for successfully sending a packet will be decreased if the 802.11 DCF is modified with the new approach of bit-free control frames. We may use $\frac{1}{r}$ ($r < 1$) to denote the improvement factor of the effectiveness of the control frames in reducing the probability of collisions caused by hidden terminals. We may also denote the length reduction factor for the control frames by v ($v < 1$). Then, the medium time needed for successfully sending a RTS frame with the modified protocol is

$$T'_{rts} = \frac{1}{(1 - p_c) \times (1 - p_h)} \times (T_{bo} + v \times t_{rts}), \quad (3)$$

and the time for successfully sending a packet in such a case is

$$T' = \frac{1}{1 - r \times f \times p_h} \times (T'_{rts} + v \times t_{cts} + t_{data}) + v \times t_{ack}. \quad (4)$$

We now show by an example how the modified protocol with bit-free control frames may reduce the control overhead and thus increase the throughput of a network. For easy reference, we named the modified MAC protocol as CSMA/FP, which denotes Carrier Sense Multiple Access with Frame Pulses (bit-free frames may be regarded as a type of in-band pulses). In the example, the network has a DSSS physical layer, the control and data frames are transmitted at 1 Mb/s and 2 Mb/s, respectively, and each packet has a size of 512 bytes. In addition, p_h assumes a value of 0.2, which means that a frame without medium reservation has a probability of 0.2 to encounter a collision caused by hidden terminals. The hidden terminal residue factor f assumes a value of 0.2 for the IEEE 802.11 DCF in the example. T_{bo} takes the value of 2 ms for a high network load case, which is a typical value shown by our simulation results in Section 4. Finally, r and v assume values of 2 and 0.4, respectively, in the example.

Fig. 2 shows the average medium time consumed for successfully delivering a packet with the two protocols in our example as the probability of a contention collision on a frame increases (i.e., as the number of nodes and/or the traffic load increase in the network¹). As shown in the figure, the performance gains of CSMA/FP over IEEE 802.11 DCF may be more than ten percent in our example.

¹ Although these factors may also affect p_h , we assign p_h a fixed value for the simplicity of demonstration.

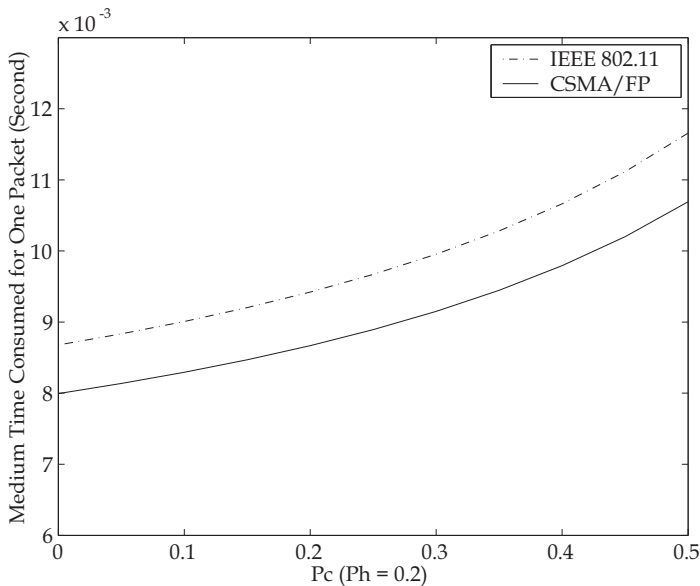


Fig. 2. Performance Analysis and Comparison

These numerical results in our example may not reflect what happens exactly in reality, since some heuristic assumptions have been made in the analysis. However, these results demonstrate the potential to considerably improve the performance of the IEEE 802.11 DCF by enhancing its capability of dealing with hidden terminals as well as shortening its control frames.

3. Applying the new approach

3.1 Basics

The challenge in applying the new approach to the IEEE 802.11 DCF is the limited capacity of the bit-free control frames in carrying control information. Particularly, only the airtime of a control frame can carry control information. To address this issue, we use two basic strategies. One is that the bit-free control frames only carry the *indispensable* information for medium access control, while the other is to use frame *pairs* for backoff duration control.

For sending bit-free control frames, we assume that the IEEE 802.11 hardware has some modification so that it can be commanded to transmit the carrier for a specified amount of time. We also assume that the airtime of a control frame can be recorded with a degree of accuracy depending on the hardware, bandwidth, and channel conditions. One protocol parameter, the minimum guard gap between the lengths of two control frames, may be adjusted based on the recording accuracy. In fact, with its carrier sense capability, the existing IEEE 802.11 hardware may record the airtime of an incoming frame.

In addition, a bit-free control frame cannot be mistaken as a bit-based frame, since a bit-free frame does not include a physical layer preamble and thus the synchronization on the frame cannot be done. A bit-based frame, however, may be mistaken as a bit-free frame if the synchronization on the frame fails. This kind of interference is usually filtered out due to the typically long airtime of a bit-based frame and the short airtime of a bit-free control frame.

3.2 Bit-free control frames

The frame type needs to be specified for each frame so that the receiver knows how to interpret the bits in the bit-based frame case or the frame airtime in the bit-free frame case. Bit-free frames carry no meaningful bits so that the frame type information can only be delivered by their airtimes. Particularly, if the airtime of a bit-free frame falls into a specified range or ranges, then the frame belongs to the type of frame denoted by the range or ranges. Besides the frame type information, the other indispensable information in a RTS frame is the address of the receiver. The length of a bit-free RTS frame needs to fall into the designated range or ranges. We therefore may not be able to encode the address information of each single receiver into the airtime of a bit-free RTS frame. To address this problem, we apply a “Mod- n ” calculation on each receiver address before it is encoded. Basically, we first divide the address by n and then encode the remainder into the frame airtime. Particularly,

$$\text{If } r = \text{Mod}(RA, n), \text{ then } F_L = \text{RTS}(r)$$

where RA is the receiver address, n is an integer, r is the remainder, F_L is the airtime of the bit-free RTS frame to send, and $\text{RTS}(r)$ is an r -indexed element in the set of RTS lengths in *microseconds*.

The Duration field in a bit-based RTS frame is also important because it specifies the period during which a receiver of the frame should back off. A bit-free RTS frame does not have the capacity for the duration information. Instead, a receiver of a bit-free RTS frame starts to back off upon receiving the frame and ends the backoff only after the medium has been sensed idle for a specified amount of time (more details later).

In our proposed design with bit-free frames, all CTS frames have the same fixed length that distinguishes them from other bit-free frames. In addition, we use control frame pairs to communicate the backoff duration information of a traditional CTS frame, which will be introduced later. Similarly, all bit-free ACK frames in our design have the same fixed length that distinguishes them from other types of bit-free frames (the address issue of these frames is discussed in Section 3.5).

In addition to the RTS, CTS, and ACK bit-free frames, we add another type of bit-free control frame named CTS-Fail frame in our design. A CTS-Fail frame has a fixed length and is sent by a CTS frame sender in two cases to notify other nodes to end their backoff. The first case is that a CTS frame sender does not receive any packet after sending the CTS frame. The second case is that a CTS frame sender receives a packet after sending the CTS frame but finds that either the packet is not intended for it or the packet has errors.

3.3 Frames working together

To explain how the four types of bit-free control frames work together in the modified IEEE 802.11 DCF, we describe how a node contends for the medium when it has a packet to transmit. The IEEE 802.11 DCF is basically a CSMA/CA protocol, and our modifications to the protocol are only on the CA part.

When a node has a packet to transmit, it starts to listen to the channel. If the channel has been found idle for a period of time longer than the DCF Interframe Space (DIFS), the node starts a random backoff timer whose value is uniformly drawn from the node’s contention window (CW). If the node detects no carrier before its backoff timer expires, it proceeds to transmit the bit-free RTS frame upon the expiration of its backoff timer. Otherwise, the node backs off.

As soon as the backoff timer of the node expires, the node starts to transmit the bit-free RTS frame. As explained earlier, the airtime of the bit-free frame is determined by the address of the intended receiver. After finishing the transmission, the node waits for a bit-free CTS frame, whose airtime is fixed and known.

After a neighbor of the initiating sender receives the bit-free RTS frame, it does the “Modn” calculation on its own address and compares the remainder to the length of the received frame in microseconds. If the remainder matches the length, the neighbor sends out a bit-free CTS frame and then waits for a packet. If the CTS frame sender does not receive any packet after a period of SIFS (Short Interframe Space) plus propagation delays, it sends out a CTS-Fail frame. On the other hand, if the remainder does not match the length of the received RTS frame, the neighbor will enter backoff and remain in backoff until the medium has been sensed idle for a period of time that is SIFS plus either the CTS frame length or the ACK frame length, whichever is longer.

After the initiating sender obtains the bit-free CTS frame, it waits for SIFS and then starts to transmit the packet. If for any reason the RTS frame sender fails to obtain the expected CTS frame, the sender starts over to contend for the medium. In such a case, the sender doubles its CW. On the other hand, if a node receives an unexpected bit-free CTS frame (i.e., the node is not a RTS frame sender), the node increases its CTS frame counter Num_{cts} by one, starts a backoff monitor timer, and then enters backoff. Such a node exits backoff in two cases. One is that its CTS frame counter Num_{cts} reaches zero when the node decrements the counter by one after receiving an ACK or CTS-Fail frame (the backoff monitor timer is canceled in such a case), while the other is that its backoff monitor timer expires (more details later).

After the initiating sender succeeds in contending for the medium, receives the expected CTS frame, and fully transmits the packet, it expects a bit-free ACK frame from the receiver. If the sender does not obtain the expected acknowledgment, it doubles its CW and starts to monitor the channel again for a retransmission.

On the other hand, after a node receives the data packet, it checks if the packet is intended for it and free of error. If so, the node sends back a bit-free ACK frame. If the packet is not intended for it or the packet has errors, the node checks whether it has sent a CTS frame for the packet. If so, the node sends out a CTS-Fail frame to notify its neighbors to exit backoff.

The whole process repeats until the initiating sender obtains an acknowledgment for the packet or the retry limit is reached. The node discards the packet in the latter case and resets its CW to the minimum size in both cases.

3.4 Some design considerations

The first design consideration on the modified MAC protocol is the choices of receive power thresholds for its bit-free control frames. Unlike bit-based frames, bit-free control frames can be correctly received as long as they can be sensed. The receive power threshold for a bit-free control frame may thus be adjusted for controlling the transmission range of the frame. As introduced earlier, a bit-based CTS frame may not successfully reach all the hidden terminals of the initiating sender (12). A node with the modified protocol, therefore, needs a lower receive power threshold for bit-free control frames.

The lowest power threshold that a node may use for receiving a bit-free control frame is the carrier sense power threshold. In such a case, a node decodes a bit-free frame if the frame can be sensed. The implementation in our simulations uses this conservative choice to

ensure the coverage of bit-free control frames. However, there is an exception. When a node receives a bit-free RTS frame matching its address, the node responds with a CTS frame only if the received power of the RTS frame is above the receive power threshold for data frames, since the node should not respond if it cannot correctly receive a packet from the other node. Another design consideration on bit-free control frames is the set of lengths in terms of airtimes that the frames should use. The basic rule is that control frames should be easy to detect and distinguish from one another. The shortest control frame in our simulations is $20\mu_s$ and the minimum guard gap between two lengths in the set is $5\mu_s$, which corresponds to 5-bit airtime at the transmission rate of 1Mb/s (even in broadband systems such as 802.11g, the control frames are recommended to be transmitted at the basic link rate). In reality, the minimum guard gap should be set based on the length detection accuracy of bit-free frames, which may be affected by the hardware, bandwidth, and channel conditions.

When choosing the length for a specific control frame that has a fixed length, we need to consider another factor. In particular, when multiple bit-free frames arrive at the same node in the same time segment, they may form a “merged” bit-free frame that has a length denoting another defined bit-free control frame. This kind of false control frame may appear when the merged frame has a longer airtime than any individual merging frame, as demonstrated by Case 3 in Fig. 3.

The possible adverse effects of the phenomenon of merged frames are alleviated by the discrete lengths of the defined control frames and the strict timelines for receiving CTS and ACK control frames. Particularly, only when a merged frame matches a defined bit-free control frame, would it possibly cause some harm. Moreover, for such control frames as CTS and ACK, a false frame may be harmful only if it emerges in the right timeline and at the right node.

However, we may still further address the merged frame phenomenon by carefully choosing the lengths for the fixed-length control frames. We have three types of fixed-length control frames, which are CTS, ACK, and CTS-Fail. Among them, a false CTS frame would arguably generate the worst scenario, in which the nodes receiving the false frame enter backoff and wait for a non-existing ACK or CTS-Fail frame for exiting backoff. Therefore, to avoid false CTS frames generated by merging frames, we need to assign a CTS frame the shortest length in the chosen length set for control frames.

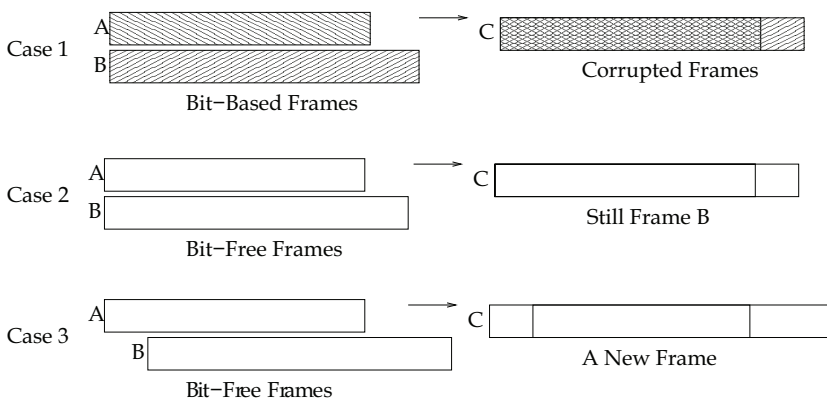


Fig. 3. Merging of Control Frames.

What happens if a false CTS frame emerges anyway due to such a reason as environmental noise? A backoff monitor timer is used to address this problem. When a node receives a CTS frame, it starts a backoff monitor timer before it enters backoff. The backoff monitor timer is set to a value T_m that is the transmission time of the largest allowable frame in the network. The node exits backoff anyway when its back off monitor timer expires. Additionally, a backoff monitor timer also solves the problem of lost ACK or CTS-Fail frames due to interference or failed nodes.

In addition, it needs some extra caution to receive a CTS frame. A RTS frame may be interpreted by two or more nodes as being intended for them due to the "Mod-n" calculation design and thus two or more bit-free CTS frames may be generated for a single RTS frame. The consequence in such a case is that the received CTS frame may be slightly longer than usual because of the various propagation delays between the RTS frame sender and its receivers (besides, the medium may be reserved in a larger space than necessary for the transmitter in such a case). A degree of tolerance on length variation is therefore needed for decoding a CTS frame. Particularly, if we denote the transmission distance of a node by d_{tx} and the signal propagation speed by c , then the decoding tolerance δ on the length of a CTS frame should be

$$\delta = 2 \times \frac{d_{tx}}{c}. \quad (5)$$

Finally, a bit-free ACK frame needs to have a longer length than a bit-free CTS-Fail frame. This is because a data frame may have one or more CTS-Fail responses besides the ACK response. In such a case, a sender still needs to recognize the ACK frame even if it is accompanied by CTS-Fail frames.

3.5 More design issues on bit-free control frames

As explained earlier, bit-free control frames have two advantages over bit-based control frames. One is that bit-free control frames are robust against interference and channel effects, and the other is that they can be very short. However, bit-free control frames have disadvantages too. One is that two or more bit-free control frames may merge at a node and form a new, false control frame. The second disadvantage of bit-free control frames is that they carry no specific address information so that they may be interpreted by any receiver as legitimate. One basic observation is that when an initiating sender is expecting a CTS or ACK frame, it has already notified its neighbors except the intended receiver to backoff. Therefore, an initiating sender may only receive a CTS or ACK frame from the intended receiver in a general case. Moreover, an initiating sender sets a strict timeline for receiving a CTS or ACK frame. For these two reasons, an initiating sender can hardly receive a false and harmful CTS or ACK frame, which makes the lack of address information in the CTS and ACK frames almost harmless. This is the reason why we choose bit-free ACK frames instead of the traditional, bit-based ACK frames in our design.

There is a special case to consider, which is that two senders may start to transmit their RTS frames almost at the same time. If the two nodes can hear each other, there is usually no harm. This is because in such a case the sender with a shorter RTS frame will finish its RTS frame transmission earlier and thus detect the other sender. If the two senders cannot hear each other, there may exist a harmful situation in which one sender overhears the CTS frame intended for the other sender and mistakenly starts to transmit its packet. This kind of

harmful situation occurs, however, with low probabilities because two senders with different RTS frames have different timelines for receiving their CTS frames.

RTS and CTS-Fail frames are more sensitive to false frames because they have no strict receive timelines. However, several factors greatly lower the possibility of harmful false RTS and CTS-Fail frames. First, neighboring nodes cannot generate false frames. Two neighboring nodes may transmit in the same time segment only if they start to transmit at almost the same time so that none of them hears the other. In such a case, the longer frame will “hide” the shorter one, as illustrated by Case 2 in Fig. 3. Second, control frames have designated lengths so that a false frame is harmful only if it has a matching length. Thirdly, not all false control frames can cause significant harm. For example, if a false RTS frame does not form at a node having a matching address, there is no harm.

In summary, the disadvantages of bit-free control frames are greatly alleviated by the following factors. First, false control frames may be few in the network because of the discrete lengths of the defined control frames. Secondly, only if false control frames form at the right node and possibly at the right time, do they cause harm. Finally, when a sender is expecting a CTS or ACK frame, its neighbors except the intended receiver have already been in backoff in general.

4. Scheme evaluations

We have done extensive simulations with ns-2 (13) to investigate the performance of the modified IEEE 802.11 DCF (named as CSMA/FP for easy reference) and compare it to the original protocol. As mentioned earlier, we have only modified the collision avoidance (CA) part of the original protocol by applying the proposed bit-free control frame approach, while other parts of the original protocol have been kept unchanged.

4.1 Configuration details

We have first evaluated CSMA/FP in a wireless LAN with saturation traffic and compared it to the original protocol. We have then used a more general scenario of a multihop ad hoc network to investigate its performance. Particularly, we evaluated the protocols from the perspective of an individual user in the ad hoc network.

From an individual user’s perspective, a network is better if the user can have statistically higher flow throughput. Although a contention-based MAC protocol may not be always fair to contending nodes in terms of one-hop, short-term throughput, the statistical rate of a random flow in the network truthfully reflects the throughput of the network, particularly when the transport layer does not apply rate control over the flows in the network, as configured in our simulations.

The ad hoc network has 100 nodes in an area of 1000 by 1000 square meters. Each node uses a transmission power of 0.2 watts, which means a carrier sense range of about 500 meters with the default power threshold settings of ns-2. The link rate of each node is 1Mb/s (a higher rate means that more bits may be transmitted in the times saved by CSMA/FP for using more effective and efficient control frames). In addition, there is a maximum of 25 Constant Bit Rate (CBR) background flows that are randomly initialized. The routing protocol used in the simulations is the Dynamic Source Routing protocol (DSR) (14).

In modifying the IEEE 802.11 DCF with the bit-free control frame approach, we have used an n of 20 in the “Mod- n ” calculation over the receiver’s address for obtaining the length of

a RTS frame. Twenty is the average number of nodes that fall into the transmission range of a node in the ad hoc network (however, we have also investigated the impact of a halved n). The elements in the length set designated for RTS frames fall into two ranges for balancing the average length of a RTS frame with the average length of other control frames. One of the ranges is from 40 to 90 μ s, while the other is from 120 to 170 μ s (with a guard gap of 5 μ s). In addition, a CTS frame, a CTS-Fail frame, and an ACK frame have fixed lengths of 20, 100, and 110 μ s, respectively.

Actually, these parameters for bit-free control frames are chosen conservatively. The accuracy of detecting the length of a frame is affected by the hardware, bandwidth, and channel conditions. If we assume a basic link rate of 1 Mb/s (control frames are recommended to be transmitted at the basic link rate in narrow-band as well as broadband 802.11 systems), then each bit of a control frame has an average transmission time of 1 μ s. The chosen parameters for the bit-free control frames are at least multiple times of this unit and are therefore safe in reality, assuming that the bits of a conventional frame can be recovered in the channel.

For other parameters, the modified protocol shares the default ns-2 configurations with the original protocol. For example, the minimum and maximum sizes of the contention window of a node are 32 and 1024 timeslots, respectively, while a timeslot is 20 μ s. In addition, the retransmission limits are 7 and 4 for a RTS frame and a longer data packet, respectively.

4.2 Wireless LANs

Fig. 4 shows the throughput of a wireless LAN versus the number of nodes in the LAN. In the simulations, every node always has packets to send (i.e., a saturation traffic scenario) and the destination of each packet is randomly selected. In addition, each packet is 512-byte long. As shown in Fig. 4, the modified protocol has a relative throughput gain of about 15% (an absolute gain of about 100 kb/s) when there are 5 nodes in the network. As the number of nodes in the network increases, the throughput gain of the modified protocol increases too. When the number of nodes in the network reaches 25, the relative gain increases to 25% (an absolute gain of 150 kb/s).

The average medium access delay for a packet in the network is shown in Fig. 5. As shown in the figure, a packet experiences less delay when the modified MAC protocol replaces the original one in the network. These results conform to the throughput results shown above. For conciseness, we only show throughput results for ad hoc networks in the following sections.

4.3 Ad Hoc networks

The multihop ad hoc network introduced earlier provides us a more general scenario to investigate the performance of the modified protocol. The nodes in the network have random waypoint movement and have a minimum and a maximum speed of 1.0 and 5.0 m/s, respectively (the average pause time is 0.5 second). In such an ad hoc network, we have examined what percentage of the packets in a test flow in the network were successfully received by the flow receiver as the network load varied.

In particular, the two protocols were tested in a series of simulations in which the rate of the background flows varied from 0.5*512 bytes/second (B/s) to 8*512 B/s with an increase factor of 100%. The test flow, however, kept its rate *constant* at 4*512 B/s to monitor the actual throughput that it could obtain in various cases of network load.

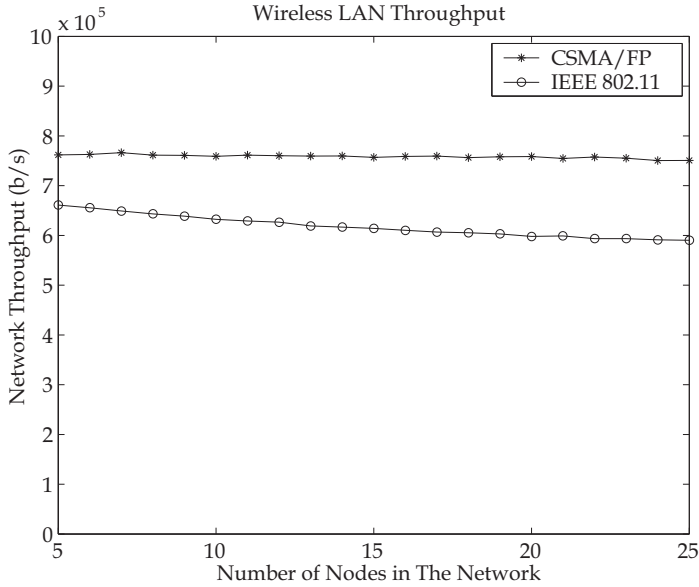


Fig. 4. Network Throughput vs. Number of Nodes

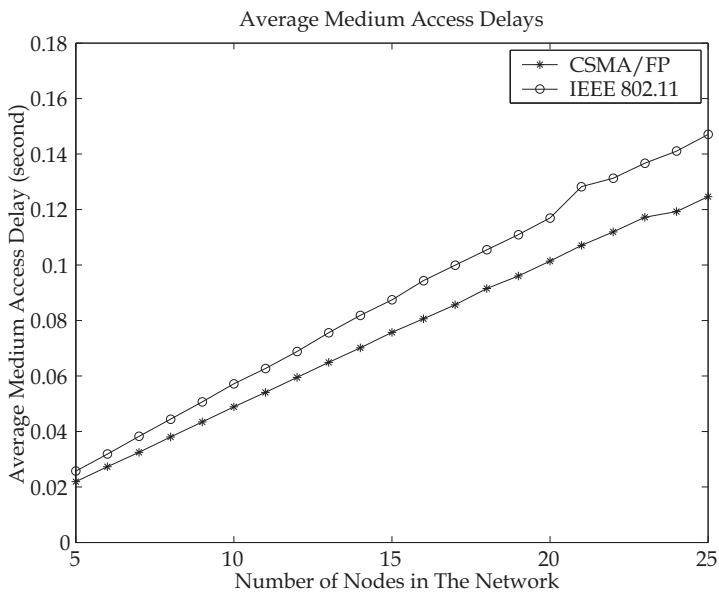


Fig. 5. Average Medium Access Delay vs. Number of Nodes

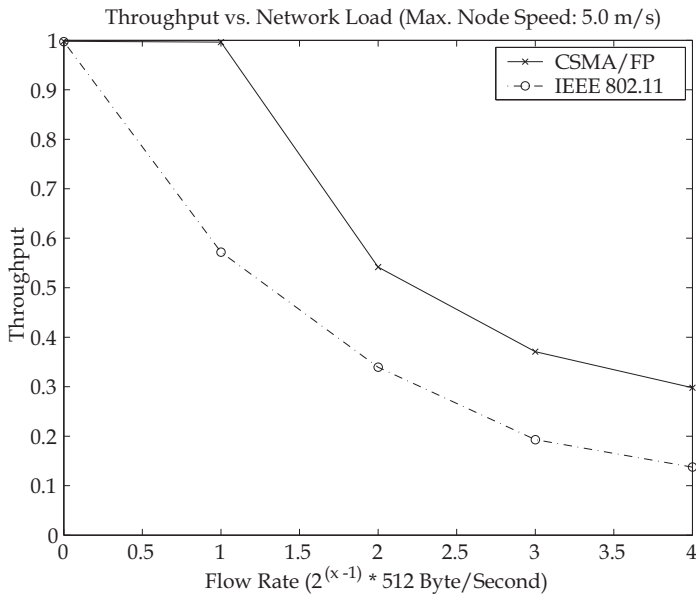


Fig. 6. Flow Throughput, Max Node Speed 5.0 m/s

Fig. 6 shows the throughput of the test flow versus the flow rate in the network, which determines the network load in our simulations. As shown in the figure, when the rate of the background flows is $0.5 \cdot 512$ B/s, almost all packets of the test flow are successfully delivered by the network with either MAC protocol. However, as the network load increases, more packets of the test flow are delivered by the network with the modified MAC protocol.

Particularly, when the rate of the background flows is $1 \cdot 512$ or $2 \cdot 512$ B/s, the throughput of the test flow increases by at least 50% as the modified MAC protocol replaces the original one. When the rate of the background flows is further increased above $4 \cdot 512$ B/s, the relative performance gains of CSMA/FP reach more than 100%. In summary, the modified protocol shows higher relative performance gains when the network load is higher.

In addition, as shown by the comparison of Fig. 6 to Fig. 4, the modified protocol shows higher performance gains in multihop ad hoc networks than in wireless LANs. These results are expected because there are hidden terminals in the multihop ad hoc network and the modified protocol is more effective in dealing with hidden terminals than the original protocol.

4.4 More hidden terminals

This section shows how the modified protocol performs when there is a higher probability of hidden terminals for a transmitter in the network. To increase the probability of hidden terminals, we increased the carrier sense (CS) power threshold of a node from less than one twentieth to half of its packet receive power threshold. The increase of the CS power threshold shrinks the carrier sense range of a node in the network.

Fig. 7 shows the throughput of the test flow when the CS power threshold has been increased in the network. As shown in Fig. 7, the relative performance gain of the modified protocol is, on average, more than 100% in the case of a higher probability of hidden

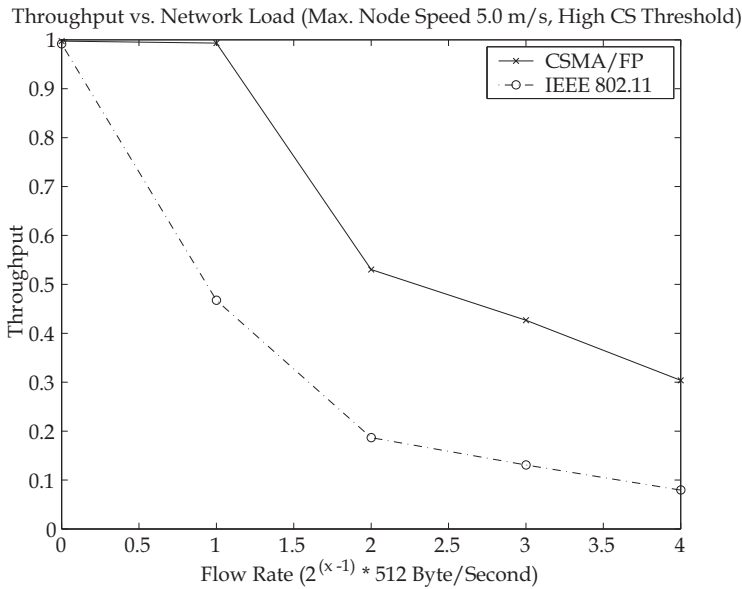


Fig. 7. Higher CS Power Threshold Case

terminals. By comparing Fig. 7 to Fig. 6, we find that the modified protocol has higher performance gains as the probability of hidden terminals is increased in the network. These results further show that the modified protocol is better in dealing with hidden terminals than the original protocol.

4.5 Rayleigh fading channel

By default, the two-ray ground channel model is used in ns-2. We have also investigated the impact of a Rayleigh fading channel on the performance of the modified protocol. The bit-free control frames of the modified protocol are robust against channel effects because of their low receive power threshold. However, a traditional, bit-based control frame may be easily lost in a fading channel.

Fig. 8 shows the results for the case of a Rayleigh fading channel. As shown by the comparison of Fig. 8 to Fig. 6, a fading channel increases the relative performance gains of the modified protocol over the original protocol. These results are expected because traditional control frames are sensitive to fading while any loss of a control frame makes all preceding related transmissions wasted.

4.6 Environmental noise

Besides the impact of channel effects, we have also investigated the impact of environmental noise on the modified protocol. On one hand, the bit-free control frames are robust against environmental noise in the sense that a noise signal may not change the length of a bit-free control frame but may corrupt a bit-based control frame. On the other hand, environmental noise may be falsely interpreted as control frames by a node with the modified MAC protocol. As explained in Section 3, a noise signal must have the right length, arrive at the right node, and possibly arrive at the right time for it to be harmful.

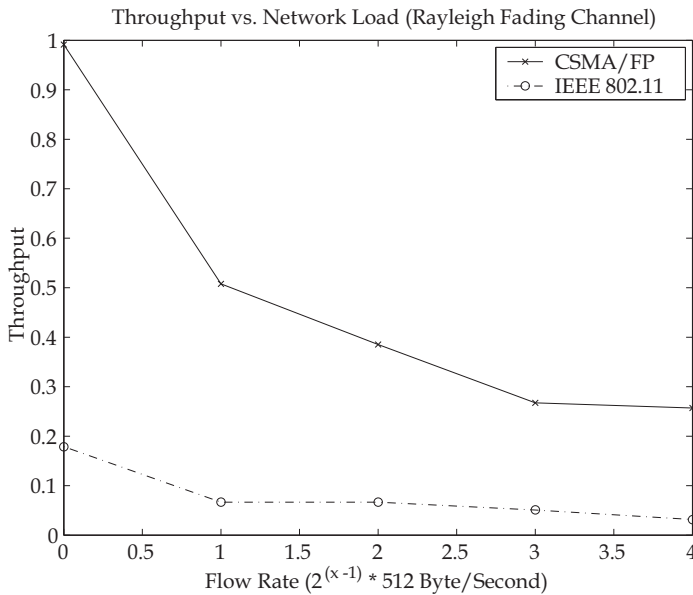


Fig. 8. Rayleigh Fading Channel Case

To test the impact of environmental noise, we placed a noise source at the center of the network and let it generate random-length noise signals at an average rate of 100 signals per second. Moreover, we restricted the noise signal lengths to the range from $1\mu\text{s}$ to $200\mu\text{s}$, which were the range designated for the bit-free control frames. The simulation results for this scenario are shown in Fig. 9. As shown by the comparison of Fig. 9 to Fig. 6, the modified protocol is not more sensitive to noise than the original one. In fact, after the noise source is introduced in the network, the modified protocol shows higher *relative* performance gains over the original one.

4.7 Protocol resilience

The above subsections are about how external factors may impact the performance of the modified protocol. This subsection shows how the parameters of the protocol affect its performance. We have investigated the three most important parameters of the protocol, which are the receive power thresholds for control frames, the length set for control frames, and the base n of the Mod- n calculations for obtaining RTS frame lengths.

Fig. 10 shows how the modified protocol performs when all its control frames use the same receive power threshold as data frames, which deprives the modified protocol of its advantage of better hidden terminal handling. As shown in the figure, the protocol still maintains significant gains over the original protocol.

Fig. 11 shows the performance of the modified protocol as the average length of its control frames becomes similar to the average length of the bit-based control frames of the original protocol. As shown in this figure, the performance of the modified protocol degrades gracefully in this case.

Fig. 12 shows how the modified protocol performs as the base n of the Mod- n calculation is halved. Halving the n is similar to doubling the node density of the network in terms of

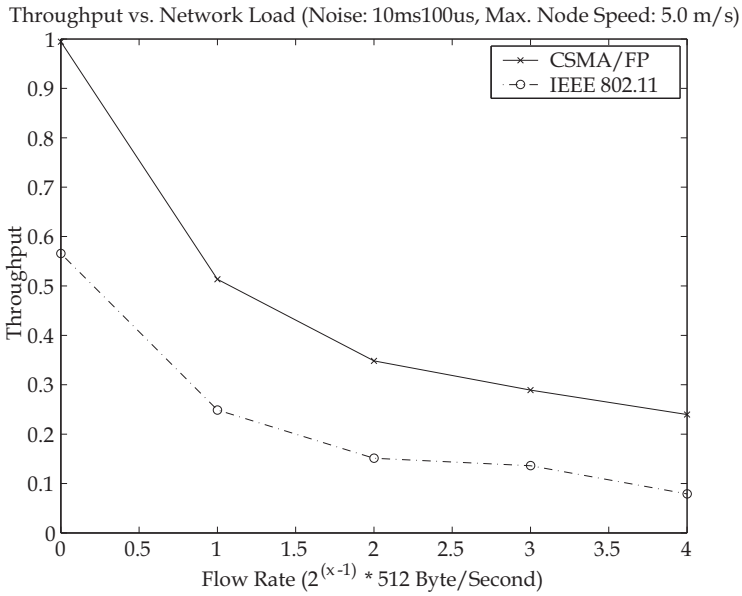


Fig. 9. Environmental Noise Case

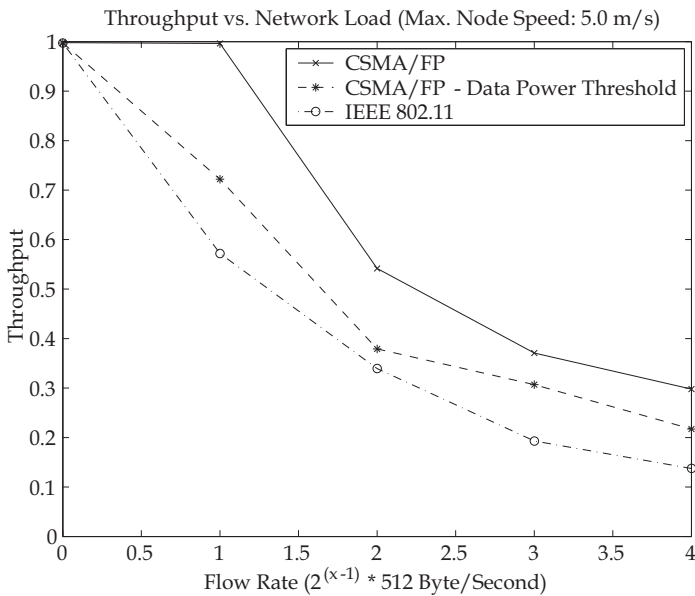


Fig. 10. Data Receive Power Threshold Case

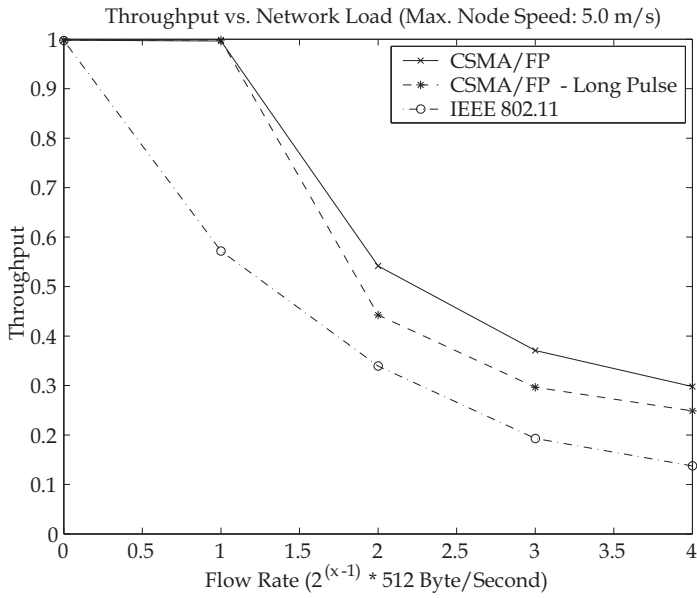


Fig. 11. Long Bit-Free Control Frames Case

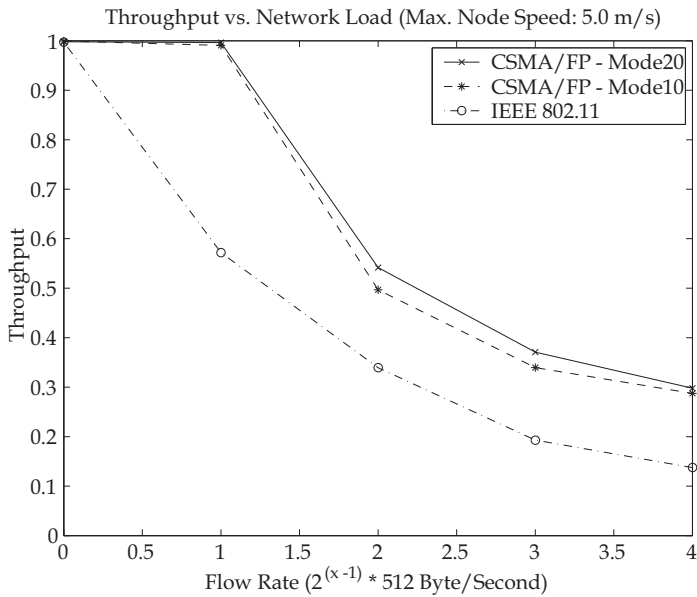


Fig. 12. Mod-n: n Changes from 20 to 10

investigating how the redundant CTS frames for a RTS frame may affect the performance of the protocol. As shown in Fig. 12, the performance of the modified protocol has a graceful degradation when the n is halved.

5. Related work

We introduce in this section some recent efforts on improving the IEEE 802.11 DCF in the community. Many efforts have been made to modify the backoff algorithm of the DCF. Cali et al. proposed an algorithm that enables each node to tune its backoff algorithm at run-time (15). Bianchi et al. proposed the use of a Kalman filter to estimate the number of active nodes in the network for dynamically adjusting the CW (16). Kwon et al. proposed a new CW adjustment algorithm that is to double the CW of any node that either experiences a collision or loses a contention (17). On the other hand, Ma et al. proposed a centralized way to dynamically adjust the backoff algorithm (18). From a theoretical perspective, Yang et al. investigated the design of backoff algorithms (19).

Another interesting scheme on backoff algorithms, named *Idle Sense*, was proposed by Heusse et al (20). With *Idle Sense*, a node monitors the number of idle timeslots between transmission attempts and then adjusts its contention window accordingly. This method uses interference-free feedback signals and the authors showed its fairness and flexibility among other features. Instead of modifying the backoff algorithm, some other works proposed diverse ways to improve the performance of the IEEE 802.11 DCF. Peng et al. proposed the use of out-of-band pulses for collision detection in distributed wireless networks (5). Sadeghi et al. proposed a multirate scheme that exploits the durations of high-quality channel conditions (21). Cesana et al. proposed the embedding of received power and interference level information in control frames for better spatial reuse of spectrum (22). Sarkar et al. proposed the combination of short packets in a flow to form large frames for reducing control and transmission overhead (23). Additionally, Zhu et al. proposed a multirate scheme that uses relay nodes in the MAC sub-layer (24).

Different from the work mentioned above, the work in this article is to improve the effectiveness and the efficiency of the collision avoidance (CA) part of the IEEE 802.11 DCF. The proposed method may work with other schemes that improve the backoff algorithm of the DCF protocol (i.e., the CSMA part of the protocol).

6. A fundamental view

Finally, we provide a fundamental view on bit-free control frames from the perspectives of information theory and digital communications. The basic goals of bit-free control frames are to increase the range, reliability, and efficiency of control information delivery for medium access control.

Information theory states that the capacity of a channel decreases as the signal to noise ratio decreases. For example, the capacity of a band-limited Gaussian channel is

$$C = W \log\left(1 + \frac{P}{N_0 W}\right) \quad (6)$$

where the noise spectral density is $N_0/2$. This equation basically states that when the received power P is lower, then the channel capacity is smaller. Therefore, if the control

information for medium access control needs to be delivered in a larger range without sacrificing reliability, then the transmission power may need to be increased (the bandwidth W is usually fixed).

There are, however, two issues with the approach of higher power for control frames. One is that the transmission power for control frames has to be increased by at least multiple times because signals deteriorate fast in wireless channels. For example, if the transmission range of a control frame needs to be doubled, then the transmission power may have to be increased by more than ten times even in free space. The other issue is that when the transmission range of a control frame is increased, then its carrier sense range is also increased at the same ratio, which causes unnecessary backoff for some nodes.

Instead, the capacity of the channel may be traded, as shown by Equation 6. The first step in this direction is to trim the control information for medium access control, which is to only deliver indispensable control information. The second step is to find away to realize the tradeoff by using new physical layer mechanisms. With bit-free control frames, the medium access control information is not translated into *bits* and then goes through the *bit* delivery process. Instead, the control information is directly modulated by the airtimes of control frames. From this perspective, the bit-free control frame approach is a cross-layer approach with which control information is delivered with a simple modulation method that trades capacity for transmission range and information reliability.

7. Conclusions

We have presented in this article a new approach of *bit-free* control frames to collision avoidance in distributed wireless packet networks. With the new approach, medium access control information is not delivered through bit flows. Instead, the information is encoded into the airtimes of bit-free control frames. Bit-free control frames are robust against channel effects and interference. Furthermore, bit-free control frames can be short because they do not include headers or preambles. We have investigated the new approach by analysis and extensive simulations. We have shown how hidden terminals, a fading channel, and environmental noise may impact the performance of the new approach. Additionally, we have examined the impact of the average length, the receive power thresholds, and the length set size of control frames on the performance of the new approach. Our conclusion is that the new bit-free control frame approach improves the throughput of a wireless LAN or ad hoc network from fifteen percent to more than one hundred percent.

8. References

- [1] F. A. Tobagi and L. Kleinrock, "Packet switching in radio channels: Part II - the hidden terminal problem in carrier sense multiple access and the busy tone solution," *IEEE Transactions on Communications*, vol. 23, pp. 1417-1433, 1975.
- [2] L. Kleinrock and F. A. Tobagi, "Packet switching in radio channels: Part i - carrier sense multiple-access modes and their throughput- delay characteristics," *IEEE Transactions on Communications*, vol. 23, pp. 1400-1416, 1975.
- [3] C. Wu and V. O. K. Li, "Receiver-initiated busy-tone multiple access in packet radio networks," in *Proc. of the ACM SIGCOMM*, Stowe, Vermont, August 1987.

- [4] Z. J. Haas and J. Deng, "Dual Busy Tone Multiple Access (DBTMA) - a multiple access control scheme for ad hoc networks," *IEEE Transactions on Communications*, vol. 50, pp. 975-985, June 2002.
- [5] J. Peng, L. Cheng, and B. Sikdar, "A new MAC protocol for wireless packet networks," in *IEEE GLOBECOM 2006*, San Francisco, CA, Nov.-Dec. 2006.
- [6] A. Colvin, "CSMA with collision avoidance," *Computer Commun.*, vol. 6, pp. 227-235, 1983.
- [7] P. Karn, "MACA - a new channel access method for packet radio," in *Proc. of the 9th ARRL Computer Networking Conference*, Ontario, Canada, 1990.
- [8] C. L. Fullmer and J. J. Garcia-Luna-Aceves, "Floor acquisition multiple access (FAMA) for packet-radio networks," in *Proc. of the ACM SIGCOMM*, September 1995.
- [9] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: a medium access protocol for wireless LANs," in *Proc. of the ACM SIGCOMM*, London, United Kingdom, August 1994.
- [10] C. L. Fullmer and J. J. Garcia-Luna-Aceves, "Solutions to hidden terminal problems in wireless networks," in *Proc. of the ACM SIGCOMM*, French Riviera, France, September 1997.
- [11] IEEE 802.11 wireless local area networks. [Online]. Available: <http://grouper.ieee.org/groups/802/11/>
- [12] K. Xu, M. Gerla, and S. Bae, "How effective is the IEEE 802.11 RTS/CTS handshake in ad hoc networks?" in *Proc. of the IEEE GLOBECOM*, Taipei, Taiwan, November 2002.
- [13] The network simulator - ns-2. [Online]. Available: <http://www.isi.edu/nsnam/ns/>
- [14] D. B. Johnson, D. A. Maltz, and Y.-C. Hu, "The dynamic source routing protocol for mobile ad hoc networks (DSR)," *IETF Internet draft, draft-ietf-manet-dsr-10.txt*, July 2004.
- [15] F. Cali, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol," *IEEE/ACM Transactions on Networking*, vol. 8, pp. 785 - 799, Dec. 2000.
- [16] G. Bianchi and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," in *Proc. of the IEEE INFOCOM*, 2003.
- [17] Y. Kwon, Y. Fang, and H. Latchman, "A novel MAC protocol with fast collision resolution for wireless LANs," in *Proc. of the IEEE INFOCOM*, 2003.
- [18] H. Ma, H. Li, P. Zhang, S. Luo, C. Yuan, and X. Li, "Dynamic optimization of IEEE 802.11 CSMA/CA based on the number of competing stations," in *Proc. of the IEEE ICC*, 2004.
- [19] Y. Yang, J. Wang, and R. Kravets, "Distributed optimal contention window control for elastic traffic in wireless LANs," in *Proc. of the IEEE INFOCOM*, 2005.
- [20] M. Heusse, F. Rousseau, R. Guillier, and A. Duda, "Idle Sense: An optimal access method for high throughput and fairness in rate diverse wireless LANs," in *Proc. of the ACM SIGCOMM*, 2005.
- [21] B. Sadeghi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic media access for multirate ad hoc networks," in *Proc. of the ACM MOBICOM*, 2002.
- [22] M. Cesana, D. Maniezzo, P. Bergamo, and M. Gerla, "Interference aware (IA) MAC: an enhancement to IEEE 802.11b DCF," in *Proc. of the VTC*, 2003.
- [23] N. Sarkar and K. Sowerby, "Buffer unit multiple access (BUMA) protocol: an enhancement to IEEE 802.11b DCF," in *Proc. of the IEEE GLOBECOM*, 2005.
- [24] H. Zhu and G. Cao, "rDCF: A Relay-enabled Medium Access Control Protocol for Wireless Ad Hoc Networks," in *Proc. of the IEEE INFOCOM*, 2005.

Secure Trust-based Cooperative Communications in Wireless Multi-hop Networks

Kun Wang, Meng Wu and Subin Shen
*Institute of IOT, Nanjing University of Posts and Telecommunications, Nanjing,
China*

1. Introduction

The word cooperate derives from the Latin words co-and operate (to work), thus it connotes the idea of “working together”. Cooperation is the strategy of a group of entities working together to achieve a common or individual goal. The main idea behind cooperation is that each cooperating entity gains by means of the unified activity. Cooperation can be seen as the action of obtaining some advantage by giving, sharing or allowing something. Cooperation is extensively applied by human beings and animals, and we would like here to map different cooperation strategies into wireless communication systems. While the term cooperation can be used to describe any relationship where all participants contribute, we tend to use it here to describe the more restrictive case in which all participants gain. If we use it in the broader sense of simply working together, it will be apparent from the context or explicitly stated. This restricted definition of cooperation contrasts with altruism, a behaviour where one of the participants does not gain from the interaction to support others (Frank & Marcos, 2006).

Cooperation has become an academic subject of intensive study in the social and biological sciences, as well as in mathematics and artificial intelligence. The most fundamental finding is that even egoists can support cooperation if necessary. In the field of information systems, some notable illustrations of this principle have recently emerged. One example is the success of open source in which thousands of people have cooperatively created a system, such as Linux. Another example is the success of eBay, which is based on a feedback system by verifying the accumulated reputations through cooperating with others in the past, making strangers mutually trust.

Recently, Wireless multi-hop networks provide yet another realm in which cooperation among large numbers of egoists can be attained, provided that the right institutional structure can be designed and implemented. Wireless communications is a rapidly emerging area of technology. Its success will depend in large measure on whether self-interested individuals can be provided a structure in which they are proper incentives to act in a cooperative mode. Cooperative techniques can be employed across different layers of a communication system and across different communication networks. The foremost premise of cooperative techniques is through cooperation, all participants engaged in cooperative communication may obtain some benefits.

An analogy between cooperation in natural and human sciences with the world of wireless communications can sometimes be established, though it is not our aim here to identify all such possibilities. It is interesting to note that in nature cooperation can take place at a small scale (i.e., few entities collaborate) or large scale (i.e., massive collaboration). The latter includes cooperation between the members of large groups up to the society itself. A similar classification holds in the wireless domain. A few nodes (e.g., terminals, base stations) can cooperate to achieve certain goals. The foreseen wireless knowledge society is expected to be a highly connected (global) network where virtually any entity (man or machine) can be wirelessly connected with each other. Cooperation in such a hyper-connected world will play a key role in shaping the technical and human perspectives of communication.

In wireless network field, Ad Hoc networking has been an attractive research community in recent years. A mobile Ad Hoc network is a group of nodes without requiring centralized administration or fixed network infrastructure, in which nodes can communicate with other nodes out of their direct transmission ranges through cooperatively forwarding packets for each other. In Ad Hoc networks, all networking functions must be performed by the nodes themselves. Each node acts not only as a terminal but also a router. Due to lack of routing infrastructure, they have to cooperate to communicate, discovering and maintain the routes to other nodes, and to forward packets to their neighbours. Cooperation at the network layer means routing (i.e., finding a path for a packet) and forwarding (i.e., relaying packets for others). While nodes are rational, their actions are strictly determined by their own interests, and each node is associated with a minimum lifetime constraint. Therefore, misbehavior exists, and it also occurs to multi-hop cellular networks. Misbehavior means deviation from regular routing and forwarding. It arises for several reasons; unintentionally when a node is faulty for the linking error or the battery exhausting. Intentional misbehavior can aim at an advantage for the misbehaving node or just constitute vandalism, such as enabling a malicious node to mount an attack or a selfish node to save energy. Malicious nodes are nodes that join the network with the intent of harming it by causing network partitions, denial of service (DoS), etc. The aim of malicious node is to maximize the damage they can cause to the network, while selfish nodes are nodes that utilize services provided by others but do not reciprocate to preserve their resources. These nodes do not have harmful intentions toward the network, though their Denial of Service actions may adversely affect the performance of the network, and turn the wireless network into an unpractical multi-hop network. The aim of selfish nodes is to maximize the benefits they can get from the network. In game-theoretic terms, cooperation in mobile ad hoc networks poses a dilemma. To save battery, bandwidth, and processing power, selfish nodes will refuse to forward packets for others. If this dominant strategy is adopted, however, the outcome isn't a functional network when multi-hop routes are needed, and all nodes are worse off. Therefore, incentive cooperation will inevitably be the key issue in cooperative communications.

In the social network, trust relationship is the essence of the interpersonal relationship. The trust among individuals depends on the recommendation of others; at the meanwhile, the credit of recommenders also determines the credit of the one they recommend. Actually, this kind of interdependent relationship composes an alleged web of trust (Caronni, 2000). In such a trust network, the trust of any individual is not absolutely reliable, but can be used as other individual's reference for their interactions. The individuals in web of trust and interpersonal network have great similarities, which are reflected in:

1. In the network, individuals in the interaction may leave sporadic "credit" information;
2. Individuals have full right to choose interactive objects;
3. Individuals have the obligation to provide recommended information to other individuals in the network.

Thus, using some conclusions from the sociological research for reference to apply all these notions to the problem of reliable packet delivery in MANETs becomes possible. However, Trust establishment is an important and challenging issue in the security of Ad Hoc networks. The lack of infrastructure in MANET makes it difficult to ensure the reliability of packet delivery over multi-hop routes in the presence of malicious nodes acting as intermediate hops.

Before we can compare different trust evaluation methods or discuss trust models for Ad Hoc networks, a fundamental question needs to be answered first. What is the physical meaning of trust in Ad Hoc networks? The answer to this question is the critical link between observations (trust evidence) and the metrics that evaluate trustworthiness. In Ad Hoc networks, trust relationship can be established in two ways. The first way is through direct observations of other nodes' behaviour, such as dropping packets etc. The second way is through recommendations from other nodes. Without clarifying the meaning of trust, trustworthiness cannot be accurately determined from observations, and the calculation/policies/rules that govern trust propagation cannot be justified.

Another security issues of distributed networks such as P2P, Ad hoc and wireless sensor networks have also drawn much attention. Cooperation between nodes in distributed networks takes significant risks, for a good node in an open network environment may suffer malicious attacks while obtaining reliable resources. Such attack can lead to the decline in the availability of network application.

Distributed trust management can effectively improve the security of distributed network. A reputation model is constructed based on the historical transactions of nodes. When a node determines to cooperate with another node, the trust value of the node should be taken into consideration first (Paola & Tamburo, 2008).

Nodes in reputation model share the result of transactions. A node considers evaluations of another node from transaction history when determining to make transactions. These evaluations may be incorrect sometimes so the research on the relationship between an evaluating node and a node being evaluated is worth exploring. It can help the reputation model decrease malicious evaluation, collect more subjective evaluations and eventually calculate the global trust value.

Current reputation models often adopt single trust, which fails to fully describe node behavior. Also, reputation model mainly researches on methods of trust measurement and analyzes the effectiveness of mathematical model with global trust value. However, the issue whether the established mathematical model is vulnerable or not is rarely discussed.

In this way, we introduce the trust model of social networks into reputation model in multi-hop networks, construct a global dual trust value for each node dramatically based on the nodes historical transactions, present a robust, cooperative trust establishment scheme in the model that enables a given node to identify other nodes in terms of how "trustworthy" they are with respect to reliable packet delivery and discuss how this model manages to resist different attacks. The proposed scheme is cooperative in that nodes exchange information in the process of computing trust metrics with respect to other nodes. On the other hand, the scheme is robust in the presence of malicious nodes that propagate different attacks.

The rest of the chapter is organized as follows: section 2 briefly introduces the related work with the writer's research and point of view, and then proposes a reputation-based trust management model in multi-hop network in section 3. Section 4 introduces an updating algorithm of trust value, so that the reputation model itself can effectively resist different attacks. Simulation results are presented in section 5 to prove the validity of the model. Section 6 discusses security issues in trust model in detail, and compares some related trust model with our research. Finally, section 7 concludes the chapter and points out some aspects of future research.

2. State-of-the-art

Cooperative techniques in wireless networks can be classified as follows (Frank & Marcos, 2006), shown in Fig. 1:

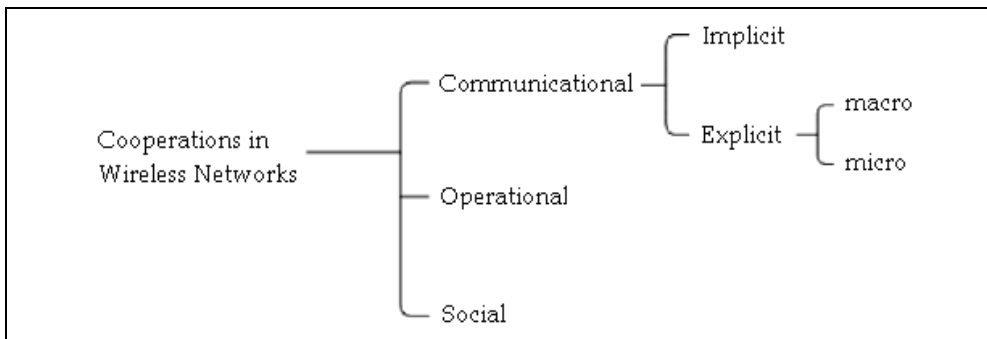


Fig. 1. A practical classification of cooperation in wireless networks

1. Communicational cooperation, which can further categorize cooperation as either Implicit, or Explicit Macro, or Explicit Micro (Functional) Cooperation. Examples of implicit cooperation are communication protocols such as TCP and ALOHA. In such protocols, participants share a common resource based on fair sharing of that resource but without the establishment of any particular framework for cooperation. In contrast, explicit macro cooperation is characterized by a specified framework and established by design. Cooperative entities that fall in this category are wireless terminals and routers, which may cooperate, for example, by employing relaying techniques that extend the range of communication for users beyond their immediate coverage area. Such cooperation potentially provides mutual benefits to all users. Explicit micro or functional cooperation is also characterized by a specific framework that is established by design. However, the cooperation involves functional parts or components of various entities, such as antennas in wireless terminals, processing units in mobile computing devices, and batteries in mobile devices. Explicit micro cooperation provides the potential for building low complexity wireless terminals with low battery consumption.

2. Operational cooperation, referring to the interaction and negotiating procedures between entities required to establish and maintain communication between different networks. The main target here is to ensure end-to-end connectivity, where the main players are (different) terminals operating in different networks. Network architecture and setup procedure are the main content of this category.

3. Social cooperation, pointing out the dynamic process of establishing and maintaining a network of collaborative nodes (e.g., wireless terminals). The process of node engagement is important as each node needs to decide on its participation in this ad hoc communication, having each decision an individual and collective impact on performance. Unlike the previous categories, in this arrangement each node is in a key position as he or she ultimately decides whether to cooperate or not. Appealing incentives need be offered to the nodes in order to encourage them to cooperate. The incentives in social cooperation are our research point.

In Ad Hoc networks, the incentive schemes can be roughly classified into reputation-based system and payment-based system. Here the latter is beyond the range of our study. In reputation-based systems, nodes observe the behaviour of other nodes and take measures, rewarding cooperative behaviours or punishing uncooperative behaviours. The typical models of this scheme include CONFIDANT (Buchegger & Le Boudec, 2002), CORE (Michiardi & Molva, 2002) and SORI (He & Wu, 2004).

CORE provides three different types of trust: subjective trust, indirect trust and functional trust. The weighted values of these three trusts are then used to determine whether to cooperate or not. CORE system allows nodes in MANET gradually to isolate malicious nodes. When the reputation assigned to a neighbour node decreases below a predefined threshold, the service provided for the misbehaving nodes will be interrupted. However, CORE system doesn't take the forged situation of indirect trust into consideration, for nodes could raise indirect trust by mutual cooperative cheating.

The goal of SORI system is to resist DoS attacks, using a similar watchdog-like mechanism to monitor. The information that reputation system maintains is the ratio of forwarded packets over sent packets. However, SORI system needs to authenticate the evaluation of reputation based on Hash function, which may naturally increase the overload of the system.

CONFIDANT is a reputation system containing monitoring, trust evaluation and trust reestablishment. This system only adopts periodic decay of trust to avoid non-cooperative behaviors without providing redemption mechanism for nodes. Yet the redemption mechanism is very important to isolated nodes, because the malicious actions of these nodes may be due to other non-malicious factors (battery energy exhausting, linking error, etc.).

Currently, the reputation models can be roughly categorized as follows:

1. Reputation models based on Public Key Infrastructure (PKI). Millan et al. adopt the approach of Cross-layer Authentication (Millan, Perez, et al., 2010), the author described the design, implementation and performance evaluation of Cross-layer. The legality of these nodes can be guaranteed by the certifications from Certificate Authority (CA). Omar et al. introduces a distributed PKI certification system based on Trust Map and Threshold Encryption (Omar, Challal, et al., 2009). Node legality is secured by Certificate Chain. However, CA will inevitably cause the problems on expansibility and invalidation of single node.

2. Reputation models based on Markov Chain. Chang et al. adopts Markov Chain to determine the trust value of the single-hop node. The node whose trust value achieves the highest will be set as the central node (Chang, Kuo, 2009). ElSalamouny et al. adopts a sort of potential Markov Chain to indicate the key behaviour of the node, and makes use of the beta probability distribution and exponential decay to evaluate the trust error (ElSalamouny, Krukow, et al., 2009). However, neither of these two reputation models involves node attacks.

3. Reputation models based on Random Probability Model [7-10], such as Power-law Distribution and Bayesian. PeerTrust (Li & Lu, 2004) controls the feedback weighting by comparing the similarity of evaluation of previous co-operator, and separates the service trust and feedback trust. But with the growth of network scale, the statistical analysis of set becomes difficult. In PowerTrust (Zhou & Hwang, 2007), there is no consideration of the malicious, selfish or strategic actions. RSFN (Saurabh, Laura, et al., 2008) adopts Bayesian Model to update the reputation with new transaction evaluation, introduces the updating algorithm between dual evaluation and zone [0,1] evaluation, and uses the algorithm to avoid bad mouthing and boost attacks during the reputation establishment process. Nevertheless, there is no further discussion regarding the effects of other types of attacks.

4. Reputation models based on fuzzy control. Ganeriwal et al. introduce a central reputation model based on a trust value pair(trust/non-trust), and set 'trust', 'non-trust', 'ignore', and 'variance' as the fuzzy controlling parameters (Victor, Cornelis, et al., 2009). However, the author doesn't consider the security issues, and the problem of CA still exists. RFSTrust (Luo, Liu, et al., 2009) is a reputation model based on fuzzy recommendation. Node trust value includes five fuzzy controlling parameters. On the security issue, the author only mentioned the selfish behaviour of nodes, but no other attacks.

5. Reputation models based on direct trust and recommendation trust [13-18]. Peng et al. adopted abnormal trust series to detect the malicious and fake recommendation, and to defend against collusion attacks (Peng, He, et al., 2008). Liu et al. proposed a two dimensional reputation model based on time and context to resist collusion attack (Liu & Issarny, 2004). Li et al. use the distance weighting-based reputation model, with Distributed Hash Table (DHT) to manage the node trust value (Li & Wang, 2009). The node trust value is evaluated according to the distance between the nodes. Sun et al. discusses the multi-defence structure reputation model based on direct and recommendation trust (Sun, Liu, et al., 2008). However, the collected trust information is not comprehensive, hence leading to the inaccuracy of trust evaluation. TrustMe (Aameek & Liu, 2003) adopt the anonymity to encourage the nodes to provide the honest information without worrying about vengeance. Two IDs are distributed to each node. One is used for transaction, and the other one is used for reputation evaluation. In addition, the model uses the central login server to distribute the unique ID to reduce the cheating and newcomer attacks. However, because reputation update and searching processes happen among nodes, dishonest evaluation of node transaction can not be prevented even though transaction certificate is required for transaction evaluation, Yu et al. introduce a dual evaluation model based on feedback trust and service trust (Jin, Gu, et al., 2007). It compares these two values to resist the malicious feedback. However, little information has been done on how to determine the consistency of these two values.

Besides, many typical reputation models have failed to consider the security issue or merely considered one or several kinds of attacks without fully analyzing the malicious, selfish and strategic behaviors. For instance, EigenTrust (Sepandar, Mario, et al., 2003) introduces a fully distributed reputation model without central login server. Nevertheless, the node ID is easy to be changed. As a result, the network is vulnerable to newcomer attack. Based on wireless sensor networks, RDATA (Ozdemir, 2008) uses different models separately to discuss the trust value of perception, routing and collection to find the malicious behaviors of each phase. TOMS (Boukerche & Ren, 2008) updates trust value based on a nonlinear algorithm, and selects trust router to exchange information in order to reduce the access of malicious

nodes to some extent. Ding et al. introduce a dynamic trust management model (Ding, Yu, et al., 2008). In a P2P file sharing application, when trust value is lower than a set threshold, a message for warning nodes malicious behaviors will be sent out to other nodes so as to control the transmission of malicious files. However, this method will be taken advantage of by some malicious nodes to defame trusted nodes.

Based on the related work above, current reputation model is mainly single trust, and doesn't consider the capability of preventing attacks. Therefore, this article introduces a trust management model based on global reputation. Meanwhile, we use the updating algorithms of trust value to comprehensively analyze the resistant mechanism of this model for different attacks.

3. Reputation-based trust management model

3.1 Model outline

In multi-hop networks, nodes provide data and service for each other, and execute distributed trust management. If logic networks are distributed, non-structural and self-organized, each node in the networks will independently determine which node it will interact with. One node can receive an evaluation after providing service to the other. Therefore, nodes reputation can be considered as the integration of evaluations from others. As a service request node, it needs reputation information from service provider. Afterwards, it selects an appropriate service provider to interact with its own strategy. As a service provider, each node expects its own trust value to be as high as possible. In this way, it can have many "customers" and benefit from the model incentive mechanism as well. However, honest nodes achieve high reputation by offering honest service, while malicious nodes gain reputation by tampering or decreasing other nodes trust value so that they obtain more chances in order to be a service provider. Undoubtedly, a good service provider only responds to the node with high trust value in terms of its own strategy. As a result, a node can get better service when it works as a request node and has high trust value.

Distributed trust management model is local recommendation-based or reputation-based. In this chapter, we focus on the latter one, i.e., when selecting a service provider, each node calculates the trust value of each response node (the trust value here is the integration of local trust and global trust). Then a node selects provider with high trust value with reference to its own strategy. In this case, malicious behaviors can be controlled to a certain extent with the increase of networks robustness.

Our aim of trust management model is that an honest node only costs little to prevent malicious behaviors. We analyze and design the model according to nodes honest behaviors, malicious attacks and multi-hop networks environment (Marti & Molina, 2006).

3.2 Model design

Most of current trust management models use dual evaluation or zone $[0, 1]$ for evaluation (Yu, Singl, et al., 2004). Dual evaluation is not subjective, but it enables node to get a high trust value by a few successful transactions, which is vulnerable to outside attacks. So our model herein uses zone $[0, 1]$ for evaluation, which enhances the pluralism of trust value and also ensures the continuity of it. We set nodes initial trust value to be 0.5, and after several transactions, the trust value of honest nodes is close to 1 while that of malicious ones will drop to less than 0.5.

There are some nodes called strategy nodes. They initially behave well and get high trust value after joining in networks. Afterwards, they start to behave maliciously, reducing QoS or providing dishonest feedback. The most common method to fight against these attacks is to implement punishment mechanism to decrease their trust value. However, some strategy nodes only offer dishonest feedback but without reducing their own QoS. If single trust is employed, the trust value of these nodes will decrease sharply and cannot show their service abilities.

In view of the situation above, we set two trust values, for each node in our model. One is service trust value (STV), providing the global trust value of the service; the other is request trust value (RTV), providing the global trust value of the evaluation. Both sides evaluate each other and update STV and RTV after each transaction. This dual trust values strategy is more flexible to fight against the attacks. We here set an example to illustrate the execution process of dual trust values in detail, shown in Fig. 2:

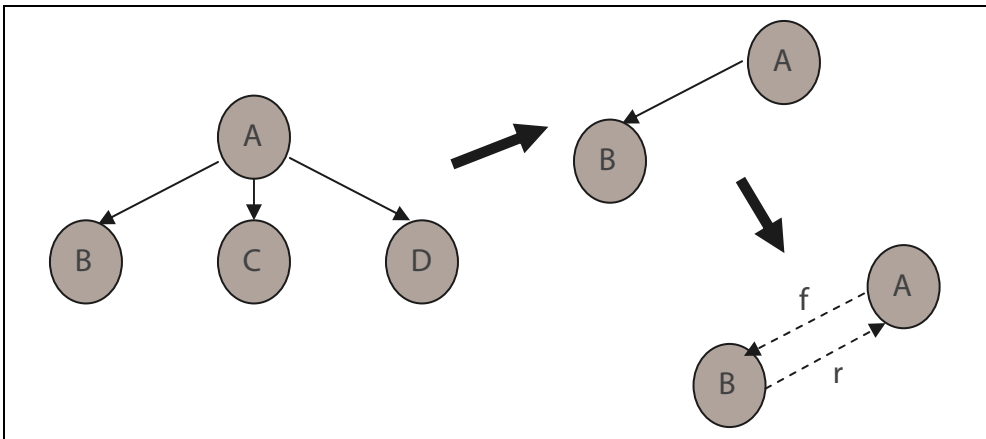


Fig. 2. Execution Process of Dual Trust Values

1. Supposing that node A has sent out a resource request and node B, C, and D have received it. They start to analyze the request and make response according to their own strategies (The analysis here includes evaluating the RTV of node A, checking whether they have such resource, etc.).
2. Node A will select the node with the highest trust value (for instance, here is node B) in terms of the local trust value (LTV: this trust value is STV stored locally, and it exists if transactions happened between them, otherwise it is set default) and the STV of responding node.
3. After selecting node B, node A will give node B an evaluation 'r' based on the transaction and its own strategies (for example, whether it is a malicious node or whether the response contains malicious information) Meanwhile, node B will give a feedback 'f' to node A as well.
4. Based on the feedback node A gives to node B, node A will calculate and update the STV of node B and save it as LTV as well.
5. Meanwhile, according to the feedback node B gives to node A, node B will calculate and update the RTV of node A.

In our model, we do not discuss which node(s) will be responsible for the calculation and storage of STV and RTV, because an agent or a neighbour node can accomplish the tasks (Thomas & Vana, 2006). To simplify the model, we suppose a central server to store and calculate STV and RTV (Zhang & Fang, 2007), while LTV is saved by a node itself.

From the view of social network, if a requester evaluates a service provider, the service provider will also evaluate the feedback of that requester. Due to revengeful psychology, feedback evaluation is normally in accord with service evaluation, that is, I will give you what you give me. Honest nodes provide honest service and feedback, while dishonest nodes provide neither honest service nor honest feedback. We can analyze the effect of mutual evaluation on reputation model by four scenarios as below, shown from Fig. 3 to Fig. 6.

Scenario 1: Service requester is an honest node while service provider is a malicious node. In this case, the mutual evaluation is bad. As a result, both the STV of malicious node and the RTV of honest node decrease. Thereby malicious nodes will have low probability to be selected as provider after some transactions.

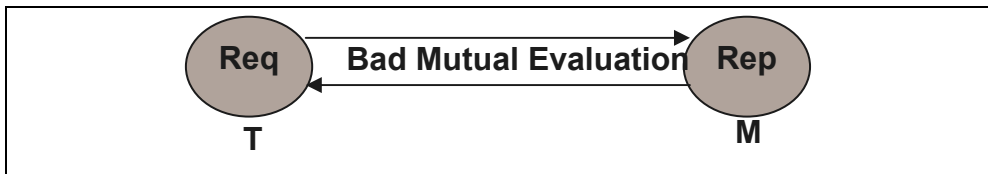


Fig. 3. Scenario 1

Scenario 2: Service requester is a malicious node while service provider is an honest node. In this case, the mutual evaluation is bad. However, service provider in our model only responds to the requester whose trust value is high. Therefore, the possibility of this scenario is very low.

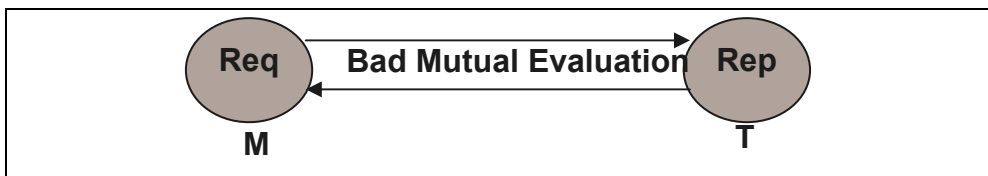


Fig. 4. Scenario 2

Scenario 3: Both service requester and provider are honest nodes. In this case, the mutual evaluation is good. When the malicious node in networks is not large scale, transactions should be in this scenario.

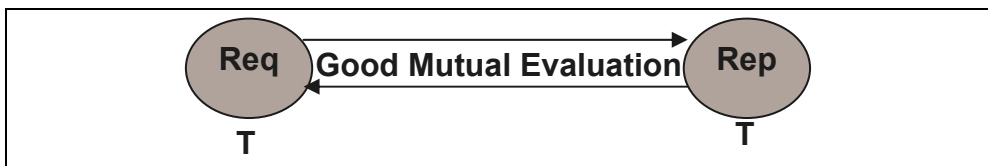


Fig. 5. Scenario 3

Scenario 4: Both service requester and provider are malicious nodes. If both sides are collusion nodes, the mutual evaluation is good. Otherwise, it is bad. The former case should be eliminated.

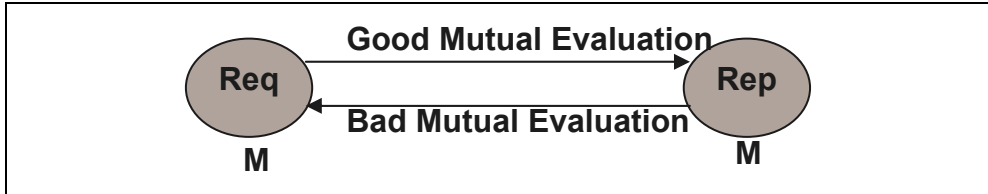


Fig. 6. Scenario 4

Some nodes only provide honest service but not honest feedback or vice versa. The influence of these nodes also includes in the scenarios above, which decreases certain trust value of the node.

Incorrect evaluation in Scenario 1, 2 and 4 will affect the reputation model. Scenario 1 and 2 can cause imputation attacks, and Scenario 4 can lead to collusion attacks. Furthermore, the effect of other attacks on the reputation model should also be considered.

4. Robustness analysis of reputation model

In this section we propose an updating algorithm of trust value to defend against different attacks. Other security problems, such as data transmission, can be resolved by encryption authentication technology. Ma et al. put forward a kind of fragment multipath transmission protocol to defend against man-in-the-middle attack, which protects the integrity of trust information (Ma & Qin, 2007). In this case, we only discuss the malicious behaviors and defending mechanism which is directly related to reputation model.

4.1 Updating algorithms of trust value

Assuming that the information collected from central node is correct and integrated, that is, not tampered or lost. We can use the statistical approach to update node STV.

Node STV updating algorithm: comparing the similarity of STV after transaction with the previous value, and updating it on the basis of original STV. More specifically, updated STV uses self-adaptive algorithm. If original STV reaches 0.5, new evaluation will affect the STV a lot. Otherwise, if STV reaches 0 or 1, the influence will be tender. Moreover, if STV this time is higher than original one, it will increase a little; otherwise, it will decrease a lot, compared with the rising extent. Here we set a mathematic model to illustrate this algorithm.

Definition: Supposing STV is T_n after the (n) th transaction of each node, and evaluation is r after the $(n+1)$ th transaction. In this algorithm, the updating STV is:

$$T_{n+1} = \begin{cases} T_n + (r - T_n) / 2 \times \theta \times \lambda_1 & r \geq T_n \\ T_n + (r - T_n) / 2 \times \theta \times \lambda_2 & r < T_n \end{cases} \quad (1)$$

Where

$$\theta = 1 - (2T_n - 1)^2 \quad (2)$$

Then define

$$\lambda = \lambda_1/\lambda_2 \quad (0 < \lambda_1 < \lambda_2 < 1) \quad (3)$$

θ in equation (1) stands for the function of T_n , shown in equation (2). It is used to adjust the weighting between historical STV and current evaluation. When historical STV is close to the original trust value, current evaluation should be emphatically considered. On the contrary, when historical trust value is far away from the original trust value, historical evaluation should be focused on.

In equation (3), λ_1 and λ_2 indicate respectively the increasing and decreasing extent of STV after each successful transaction, while λ shows the ratio of increasing extent to decreasing extent. Generally, decreasing extent is greater than increasing extent, controlled by λ . Furthermore, the algorithm can also be designed as: when the STV decreases to a given threshold even though the later transaction is honest, the increasing extent of STV will still be less than the decreasing one. In this way, the malicious behaviors can be published dramatically.

Since updating algorithms of RTV and STV are the same, there is no further discussion here.

4.2 Attacks and defenses

Strategy node attacks

Generally, strategy nodes achieve high trust value through some small transactions. Afterwards, they execute a big cheat. Repeatedly, they obtain the maximum benefits with minimum cost. Therefore, our reputation model needs to make the rising of trust value be slower. Specifically, we can adjust the value of λ in equation (3). Decreasing the value of λ , which means that the decreasing extent of trust value is much larger than the increasing extent, can prevent strategy nodes from gaining benefits.

Imputation attacks and boosts attacks

If some honest nodes are slandered by some malicious ones, their STV will decrease to quite a low level. Therefore, reputation model should be able to offer nodes the chance to regain their STV. Meanwhile, the request trust value of malicious nodes will decrease accordingly. In this way, when request trust value of malicious nodes decreases to some threshold, there are no nodes which would like to respond to them in networks. In this way, the requests of malicious nodes will be constrained, and malicious nodes would not dare to defame other nodes.

It's not comprehensive to accumulate the trust value based on only a few transactions with nodes, because boost attacks will be easy to come up among a few nodes. Therefore trust evaluation should be collected deep and extensively (Wang, Mokhta, et al., 2008). According to this idea, if transaction successful times reach out to a given value, the STV updating algorithm will change, that is, increasing extent will be slower. Specifically, we can change the λ_1 in equation (1) into $\lambda_1 \times 1/n$ (n stands for transaction successful times between two nodes) to control the deep collection of trust information.

Collusion attacks

Collusive nodes always cooperate with each other (for instance, virtual transactions) to increase their trust value, and organize together to slander other nodes with higher STV. This kind of "teamwork" attack is more harmful than single imputation or boost attack. However, since what we discuss in this chapter is logic network where information is transmitted in flooding; login server automatically creates logical neighbours for new-

joining nodes and randomly distributes them to other nodes as neighbours, a collusion group is hard to form between nodes, which can restrict collusion attacks to some extent.

Sybil attacks (Douceur & Donath, 2002) and Newcomer attacks (Resnick & Zeckhauser, 2000)

A malicious node can make Sybil attacks to reputation model by pretending to be different nodes in networks with different IDs each time. In this way, different IDs can share the decreasing of trust value so that a single ID of malicious node suffers less punishment.

If a malicious node can easily join in a network as a fresh one, it will delete its bad trust records by frequently entering and leaving the network. This is so-called Newcomer attacks.

Two schemes as below can resolve these two kinds of attacks.

Scheme 1: Login server needs some evidences to ensure that each node has one system ID. To keep login server from being open to attacks, such as DoS attack, the function of login server should be decentralized. However, fewer users would like to login if authentic and sensitive information is required. If we just bind the IP address with node ID instead of using login server, sybil attacks will be hard to fight against. SybilGuard (Yu, Kaminsky, et al., 2008) can be seen as a reference, for Yu et al. have proposed an effective protocol to wipe off "attack edge".

Scheme 2: This chapter mainly focuses on reputation model, not only encouraging new-joining nodes but also preventing newcomer attacks. Therefore, we can adopt a mechanism that nodes trust value can slowly reach a certain level which is not too high, if they succeed in previous transactions with a given number. When malicious nodes find it difficult to gain as much benefit as new-joining ones, newcomer attacks will be reduced. For example, a node trust value can finally reach the highest trust value 0.6 after previous 10 successful transactions, while the trust value is easy to drop once nodes process malicious behaviour.

Free riding attacks

There are some nodes referred as free riders in networks. They only receive the service provided by other nodes, but are not willing to provide any service or trust evaluation for other nodes. In our reputation model, the service and request trust value of these nodes all maintain at an initial level, so they can only get very limited resources. In addition, the incentive mechanism (for example, they can obtain the priority of network resources if STV reaches out to some extent) can be adopted in this model to motivate free riders to provide service and trust evaluation.

5. Performance evaluation

To verify the effectiveness of our reputation model, we presented a Java-based simulation program. We firstly checked whether the updating algorithm of trust value can control the STV of strategy nodes. Subsequently, we compared the dual trust values of nodes with expected values when malicious nodes existed in networks. At last, we analyzed how our model resisted the boost attacks.

In simulation environment, we adopted Gnutella routing architecture, with a central server storing trust value. There are totally 1000 nodes and 100 resources in simulation network and each node randomly chooses at most 5 nodes as neighbors and obtains 5 resources. The proportion of malicious nodes is not more than 50%. Averagely, each node sends 100 requests and the Time To Live (TTL) of resource request is 3. We assume that malicious nodes are always the most active ones to respond to any resource request messages. Besides, we also suppose that honest nodes provide honest service and evaluations while malicious

nodes provide fake resources and evaluations. The other common parameters for all the simulation are listed in Table 1. The results are the mean value from several simulations, demonstrated from Fig. 7 to Fig. 12.

Parameters	Description	Default
λ	Ratio of increasing extent to decreasing extent	1/8
λ_1	Trust value increasing extent	0.1
λ_2	Trust value decreasing extent	0.8

Table 1. Simulation Settings

Experiment 1: When all the nodes in network are honest nodes except one strategy node, we observed the change of STV of that strategy node, as is shown in Fig. 7.

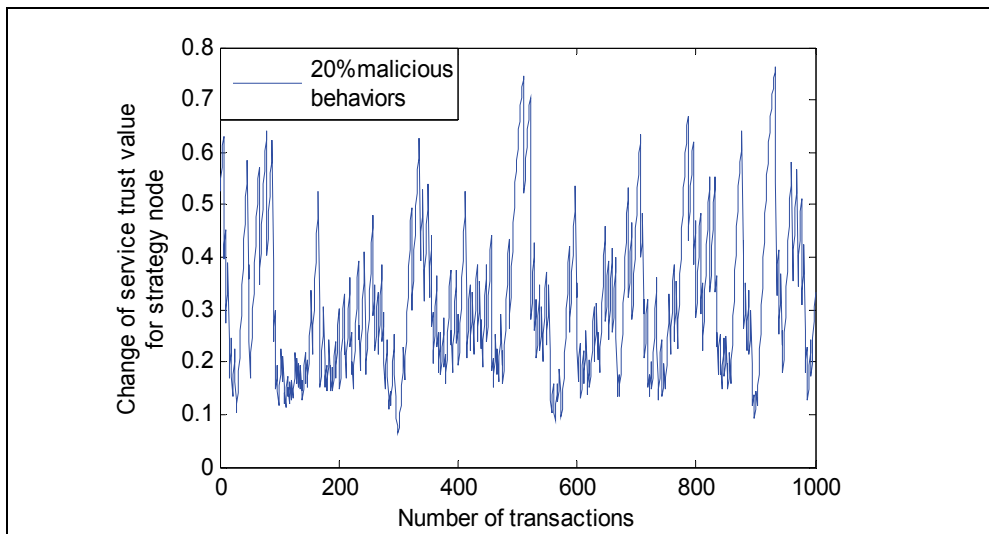


Fig. 7. Changes of STV for Strategy Nodes

In Fig. 7, when strategy node had 20% malicious service, the trust value of that node was controlled at 0.3, at most 0.75. In this case, the strategy node failed to be trusted by resource requester, which made less effect of malicious service on network.

Experiment 2: We set the network with about 30% malicious nodes and 70% honest nodes, and defined that no request could be provided when the RTV of nodes was less than 0.2. Afterwards, we run the updating algorithm of trust value to update dual trust value. After 100000 transactions, we checked whether the STV and RTV of both malicious and honest nodes were within the expected range. The results are presented from Fig. 8 to 10.

Fig. 8 indicates all the STVs of malicious nodes decreased to less than 0.5, which means honest nodes no longer chose these malicious nodes as cooperators. In Fig. 9, all the RTVs of malicious nodes reached 0.195, which was less than 0.2. In this way, these malicious nodes failed to request service. From Fig. 10 we can see that the STV of a minority of honest nodes dropped to 0.5 or less due to imputation attacks from malicious nodes. However, most of

honest nodes became trustful nodes, whose STVs approached 1. For those honest nodes whose trust values decreased, our reputation model allowed them to regain the opportunities to be trusted by offering some new services. Fig. 11 shows that the RTV of honest nodes were all more than 0.5, which means that malicious nodes, as responding nodes, could be controlled after a few transactions, and could not be selected by honest nodes again. Thus the effect was less on the RTV of honest nodes.

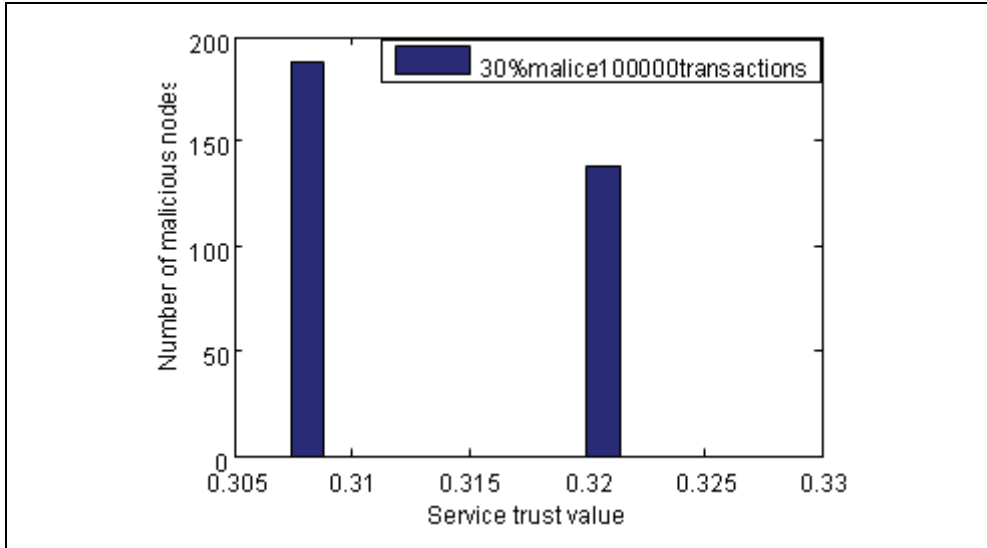


Fig. 8. STV Distribution of Malicious Nodes

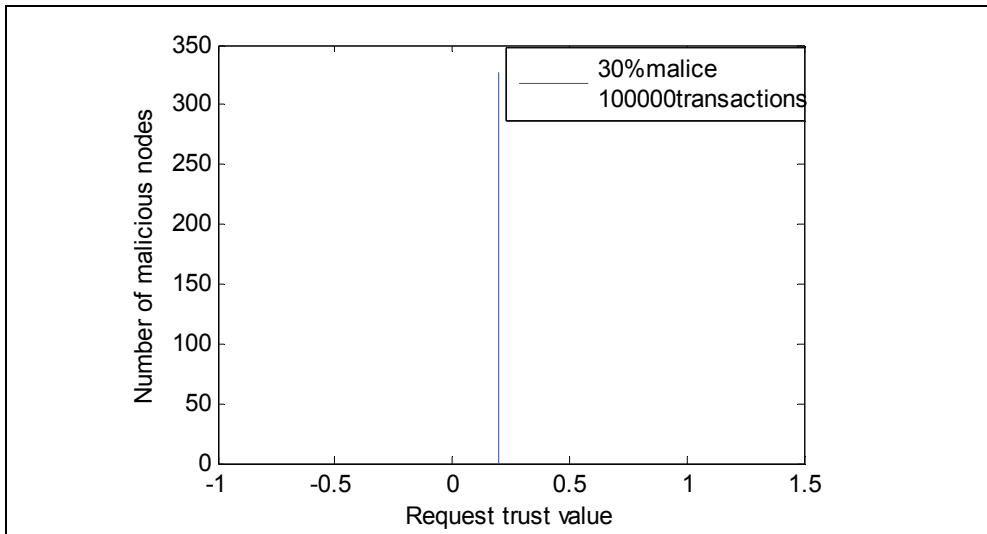


Fig. 9. RTV Distribution of Malicious Nodes

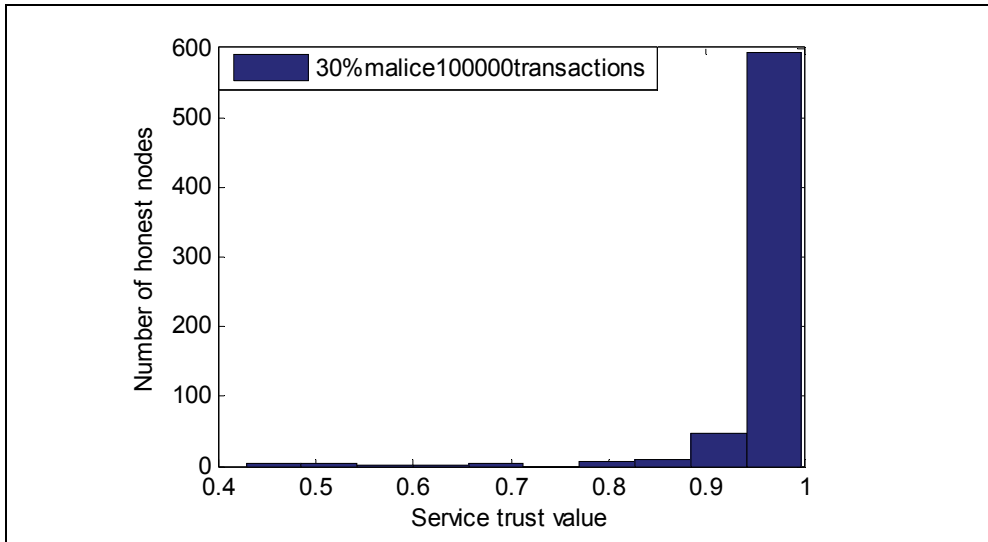


Fig. 10. STV Distribution of Honest Nodes

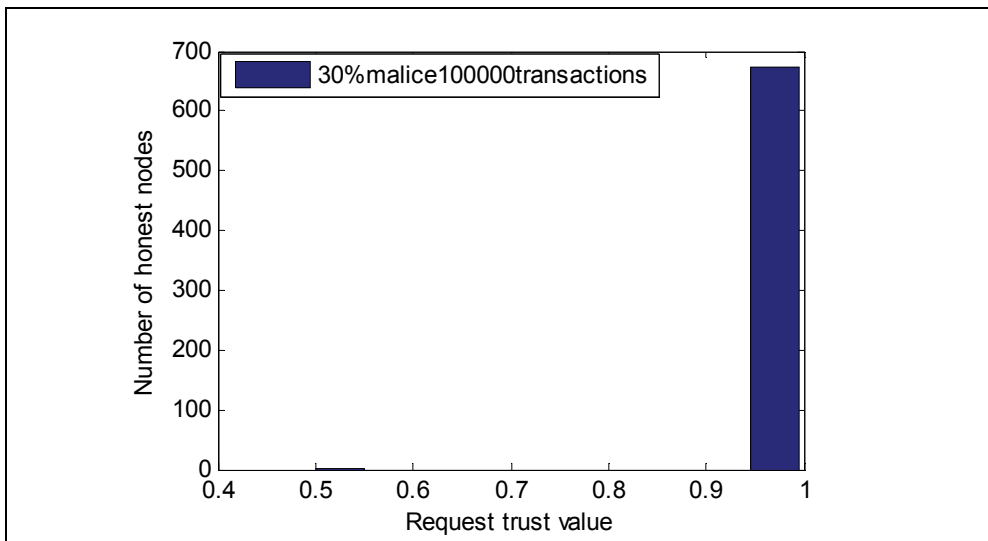


Fig. 11. RTV Distribution of Honest Nodes

Experiment 3: Boost attacks among the nodes in the network will degrade the network performance, causing more damages when happen among malicious nodes. Furthermore, boost attacks can make the STV of malicious nodes rise, hence deceiving honest nodes to transact with them. To avoid this attack, we adopted the updating algorithm with changing λ_1 into $\lambda_1 \times 1/n$. After experiments, we analyzed the changes of STV of some nodes in three conditions: non-boost, 80% boost and 100% boost, as shown in Fig. 12.

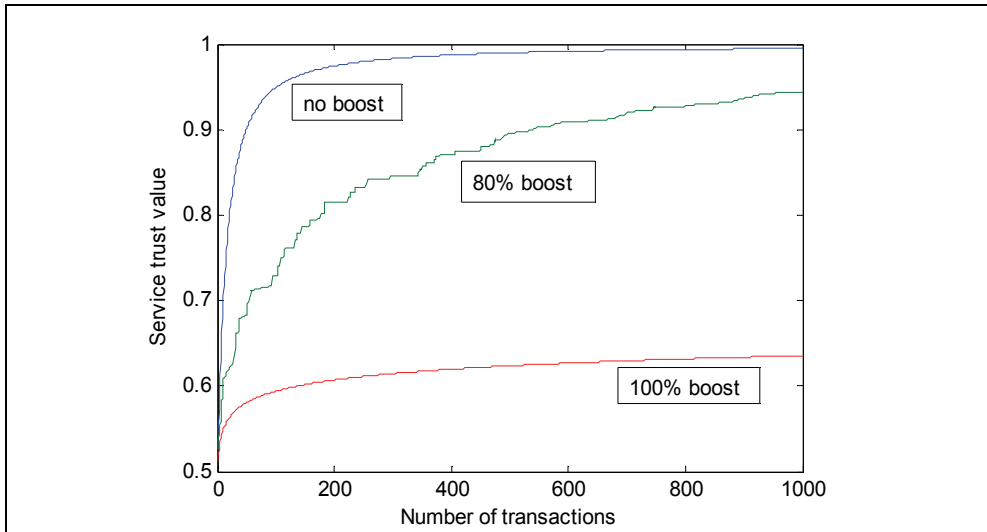


Fig. 12. Comparison of STV for Boost and Non-boost Cases

Fig. 12 illustrates the STV of that node rose slowly in condition of 80% boost, compared with the condition of non-boost. If requesting the same service, this node was selected as service provider. However, due to some honest transactions, the STV can also reach relatively high level after 1000 transactions. On the other hand, we can also see that the STV of malicious nodes in 100% boost achieved 0.6 or so. Once this node responded to honest nodes and is selected as service provider, its STV dropped to less than 0.5 right away. The comparison among these three conditions indicates that our model can effectively resist boost attacks to a certain extent.

6. Discussions and future work

Recommendation mechanism is an important component in any trust evaluation systems. The effectiveness of recommendation is closely related with communication overhead. For example, at the beginning of trust evaluation when few interactions have taken place in the network, higher mobility requires higher overhead. Then, after the trust evaluation system has been running for a long time, a mobile node has had opportunities to interact with many other nodes. Compared with a stationary node, a mobile node has a larger probability to interact with recommenders. In this case, the overhead of requesting recommendations for a node with high mobility can be reduced.

On the other hand, since trust evaluation can effectively improve network performance and detect malicious nodes, trust evaluation itself is an attractive target for attackers (Marmol & Penez, 2009).

A well-known attack is bad-mouthing attack (Dellarocas, 2000), that is, malicious parties providing dishonest recommendations to frame up good parties and/or boost trust values of malicious peers. The defense against the bad-mouthing attack has been considered in the design of the proposed trust evaluation system. First, the action trust and the recommendation trust records are maintained separately. Only the nodes who have provided good recommendations previously can earn high trust. Second, according to the

necessary conditions of trust propagation, only the trust from the entities with positive trust can propagate. Third, the fundamental axioms limit the power of the entities with low trust. Trust evaluation may also be vulnerable to the Sybil attack and the newcomer attack. If a malicious node can create several faked IDs, the trust evaluation system suffers from the Sybil attack. Here, the faked IDs can share or even take the blame, which otherwise should be given to the malicious node. If a malicious node can easily register as a new user, the trust evaluation suffers from the newcomer attack. Here, malicious nodes can easily remove their bad history by registering as a new user. The defense against the Sybil attack and newcomer attack does not rely on the design of trust evaluation system, but the authentication and access control mechanisms, which make registering a new ID or a faked ID difficult.

In terms of security issues in trust model, some literatures related to our work have done a lot researches in trust mechanism. Here I will compare some typical models with our work in detail.

S. F. Peng, et. al. showed a weighted trust formula and presented an integrated trust update method. They used abnormal trust value sequence and statistical analysis to detect malicious recommenders and false recommendation trust values in the whole lifetime of trust. In addition, trust value update analysis aims at protecting against untrue recommendation, such as the collusion problem. However, authors adopted a common formula $T(n, m) = \alpha \times T_d + (1 - \alpha) T_r$ to do the trust evaluation, which is different from our model. Meanwhile, the update of α was not considered in the paper, so we do not know the impact of α on trust formalization when it changes.

P. Victor, et. al. advocated the use of a centralized trust model in which trust scores are (trust, distrust)couples, drawn from a bilattice that preserves valuable trust provenance information including gradual trust, distrust, ignorance, and inconsistency. Authors presented a collection of four operators simultaneously in one model, especially the distrust information, which is their novel contribution. However, proposed trust techniques require a central authority to propagate and aggregate trust values; however, as the amount of nodes continues to grow, it will get more and more difficult to manage all trust information in one place, so a decentralized approach may be more appropriate. Furthermore, privacy of data is becoming increasingly important in applications, and nodes may refuse to disclose their personal trust. Authors didn't discuss the security issues in the paper.

J. H. Luo, et. al. promoted RFSTrust, a trust model based on fuzzy trust similarity to quantify and to evaluate the trustworthiness of nodes, which includes five types of fuzzy trust relationships based on the fuzzy relation theory and a mathematical description for MANETs. RFSTrust has some identification and containment capability in synergies cheating, promotes data packets forwarding between nodes, and improves the performance of the entire MANETs. But authors discuss only one type of situation when selfish nodes attack. No other types of nodes attacks are considered.

J. S. Liu and V. Issarny presented a reputation model, which incorporates two essential dimensions, time and context, along with mechanisms supporting reputation formation, evolution and propagation. Their model shows effectiveness in distinguishing truth-telling and lying agents, obtaining true reputation of an agent, and ensuring reliability against attacks of defame and collusion. The common ground of their work and ours is that we both take the time dimension of trust update into consideration, while the main difference is that Liu regards node's new behavior as a part of trust value, while we see the service and request from other nodes as a key proportion of trust establishment.

X. M. Li, et. al. gave us a global trust model, which is based on the distance-weighted recommendations under P2P circumstance (Li & Wang, 2009). Their model uses distributed

methods to quantify and evaluate the credibility of peers to identify and restrain some common collective cheatings. The global credibility relies on distance between nodes in their model, and Distributed Hash Table (DHT) is used to designate peers to manage the credibility. That is to say, hash function needs using in their model, which is different from ours. Besides these known attacks in the literatures, a malicious node may also reduce the effectiveness of trust evaluation through other methods. While the focus of this chapter is to lay the foundation of trust evaluation with meaningful trust metrics, we do not investigate all possible attacks in this chapter. Therefore, more security issues of cooperative communication in Ad Hoc networks are our following targets.

7. Conclusion

The creditability of multi-hop networks can achieve from the aspects of authentication, authorization, access control, as well as the reputation-based trust management model (Wang & Lin, 2008). In this chapter we propose a reputation model based on global STV and RTV. Afterwards, we discuss five attacks on the model in the progress of establishing reputation, and testify the model robustness of anti-attacks by simulations. Since the proposed model takes no consideration of the location issue of nodes for computing and storing trust information, our model can be applied in both structured multi-hop networks and unstructured ones.

8. Acknowledgements

We acknowledge a financial support from Six Talented Eminence Foundation of Jiangsu Province(06-E-043); Scientific Research Foundation of NJUPT (NY209016);

9. References

- Aameek, S. & Liu, L. (2003). TrustMe: anonymous management of trust relationships in decentralized P2P systems, *Proceedings of IEEE 3rd International Conference on Peer-to-Peer Computing*, pp. 142-149, ISBN: 0-7695-2023-5, Sweden, September 2003, IEEE Press, Linkoping
- Boukerche, A. & Ren, Y. L. (2008). A trust-based security system for ubiquitous and pervasive computing environments. *Computer Communications*, Vol. 31, No. 18, page numbers (4343-4351), ISSN: 0140-3664
- Buchegger, S. & LeBoudec, J. Y. (2002). Performance Analysis of the CONFIDANT Protocol (Cooperation of Nodes-Fairness in Dynamic Ad-Hoc NeTworks, *Proceedings of the 3rd ACM International Symposium of Mobile MANET Networking and Computing*, pp. 80-91, ISBN: 1-58113-501-7, Switzerland, June 2002, ACM Press, Lausanne
- Caronni, G. (2000). Walking the Web of trust, *Proceedings of the IEEE 9th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pp. 153-159, ISBN: 0-7695-0798-0, USA, June 2000, IEEE Press, MD
- Chang, B. J. & Kuo, S. L. (2009). Markov Chain Trust Model for Trust-Value Analysis and Key Management in Distributed Multicast MANETs. *IEEE Transactions on Vehicular Technology*, Vol. 58, No. 4, page numbers (1846-1863), ISSN: 0018-9545
- Dellarocas, C. (2000). Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems, *Proceedings of the International Conference on Intelligent Systems*, pp. 520-525, ISBN: ICIS2000-X, Australia, December 2000, AIS Press, Brisbane, Queensland

- Ding, X. H. ; Yu, W. & Pan, Y. (2008). A Dynamic Trust Management Scheme to Mitigate Malware Proliferation in P2P Networks, *Proceedings of IEEE International Conference on Communication*, pp. 1605-1609, ISBN: 978-1-4244-2075-9, China, May 2008, IEEE Press, Beijing
- Douceur, J. R. & Donath, J. S. (2002). The sybil attack, *Proceedings of the first International Workshop on Peer-to-Peer systems*, pp. 251-260, ISBN: 3-540-44179-4, USA, March 2002, Springer-Verlag Press, MIT Faculty Club, Cambridge, MA
- ElSalamouny, E. ; Krukow, K. & Sassone, V. (2009). An analysis of the exponential decay principle in probabilistic trust models. *Theoretical Computer Science*, Vol. 410, No. 41, page numbers (4067-4084), ISSN: 0304-3975
- Frank, H. P. F. & Marcos, D. K. (2006). Cooperation in Nature and Wireless Communications, In: *Cooperation in Wireless Networks: Principles and Applications*, Frank, H. P. F. & Marcos, D. K. (Eds.), page numbers (1-27), Springer Press, ISBN: 978-1-4020-4710-7, Netherlands
- He, Q. & Wu, D. (2004). SORI: A Secure and Objective Reputation-Based Incentive Scheme for Ad Hoc Networks, *Proceedings of IEEE Conference on Wireless Communications and Networking*, pp. 825-830, ISBN: 0-7803-8344-3, USA, March 2004, IEEE Press, Atlanta, GA
- Jin, Y. ; G, Z. M. & B, Z. J. (2007). Restraining False Feedbacks in Peer-to-Peer Reputation Systems, *Proceedings of International Conference on Semantic Computing*, pp. 304-312, ISBN: 0-7695-2997-6, USA, September 2007, IEEE Press, CA
- Kamvar, S. D. ; Mario, T. S. & Molina, H. G. (2003). The EigenTrust Algorithm for Reputation Management in P2P Networks, *Proceedings of the 12th International Conference on World Wide Web*, pp. 640-651, ISBN: 1-58113-680-3, Hungary, May 2003, ACM Press, Budapest
- Li, X. & Liu, L. (2004). PeerTrust: Supporting Reputation-Based Trust for Peer-to-Peer Electronic Communities. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 7, page numbers (843-857), ISSN: 1041-4347
- Li, X. M. & Wang, J. K. (2009). A Global Trust Model of P2P Network Based on Distance-Weighted recommendation, *Proceedings of 2009 IEEE International Conference on Networking, Architecture, and Storage*, pp. 281-284, ISBN: 978-0-7695-3741-2, China, July 2009, IEEE Press, Zhang Jia Jie, Hunan
- Liu, J. S. & Issarny, V. (2004). Enhanced Reputation Mechanism for Mobile Ad-hoc Networks, *Proceedings of 2nd International conference on Trust Management*, pp. 48-62, ISBN: 3-540-21312-0, UK, March 2004, Springer Press, Oxford
- Luo, J. ; Liu, H. X. & Fan, M. Y. (2009). A trust model based on fuzzy recommendation for mobile ad-hoc networks. *Computer Networks*, Vol. 53, No. 14, page numbers (2396-2407), ISSN: 1389-1286
- Ma, X. X. & Qin, Z. G. (2008). Partition and multi-path transmission: An encryption-free reputation sharing protocol in Gnutella-like peer-to-peer network. *Computer Communications*, Vol. 31, No. 14, page numbers (3059-3063), ISSN: 0140-3664
- Marmol, F. G. & Perez, G. M. (2009). Security threats scenarios in trust and reputation models for distributed systems. *Computer & Security*, Vol. 28, No. 7, page numbers (545-556), ISSN: 0167-4048
- Marti, S. & Molina, H. G. (2006). Taxonomy of Trust: Categorizing P2P Reputation Systems. *Computer Networks*, Vol. 50, No. 4, page numbers (472-484), ISSN: 1389-1286
- Michiardi, P. & Molva, R. (2002). Core: A COLlaborative REputation mechanism to enforce node cooperation in Mobile Ad Hoc Networks, *Proceedings of IFIP - Communication*

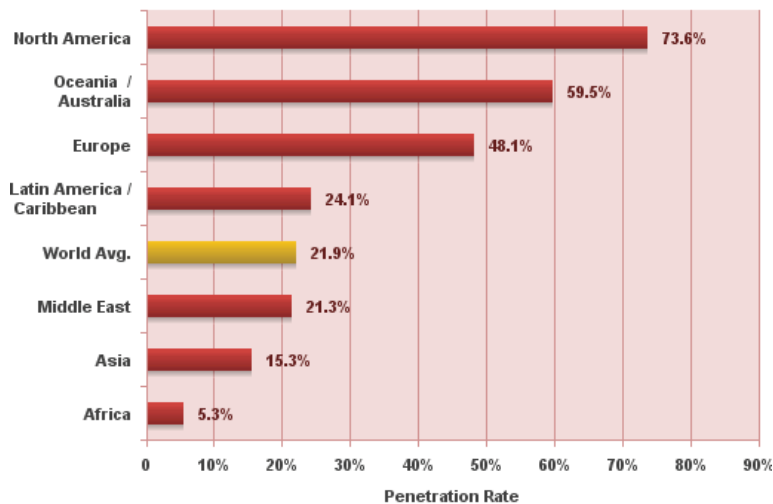
- and Multimedia Security*, pp.107-121, ISBN: 1-4020-7206-6, Slovenia, September 2002, Kluwer Press, Portoroz
- Millan, G. L.; Perez, M. G.; Perez, G. M. & Skarmeta, A. F. G. (2010). PKI-Based Trust Management in Inter-Domain Scenarios. *Computers & Security*, Vol. 29, No. 2, page numbers (278-290), ISSN: 0167-4048
- Omar, M.; Challal, Y. & Bouabdallah, A. (2009). Reliable and fully distributed trust model for mobile ad hoc networks. *Computers & Security*, Vol. 28, No. 3, page numbers (199-214), ISSN: 0167-4048
- Ozdemir, S. (2008). Functional reputation based reliable data aggregation and transmission for wireless sensor networks. *Computer Communications*, Vol. 31, No. 17, page numbers (3941-3953), ISSN: 0140-3664
- Paola, D. & Tamburo, A. (2008). Reputation Management in Distributed Systems. *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing*, pp. 666-670, ISBN: 978-0-7695-3258-5, March 2008, Malta
- Peng, S. F.; He, J. S. & Meng Y. (2008). Reputation-based Trust update in network environment, *Proceedings of 2008 International Symposium on Electronic Commerce and Security*, pp. 118-123, ISBN: 978-0-7695-3258-5, China, August 2008, IEEE Press, Guangzhou
- Resnick, P. & Zeckhauser, R. (2000). Reputation System. *Communications of the ACM*, Vol. 43, No. 12, page numbers (45-48), ISSN: 0001-0782
- Saurabh, G.; Laura, K. B. & Mani, B. S. (2008). Reputation-based framework for high integrity sensor networks. *ACM Security for Ad-hoc and Sensor Networks*, Vol. 4, No. 3, page numbers (1-17), ISSN: 1550-4859
- Sun, Y.; Zhu H. & Liu, K. J. R. (2008). Defense of Trust Management Vulnerabilities in Distributed Networks. *IEEE Communications Magazine*, Vol. 46, No. 2, page numbers (112-119), ISSN: 0163-6804
- Thomas, R. & Vana, K. (2006). Decentralized trust management for ad-hoc peer-to-peer networks, *Proceedings of the 4th international workshop on Middleware for Pervasive and Ad-Hoc Computing*, pp. 6, ISBN: 1-59593-421-9, Australia, November 2006, ACM Press, Melbourne
- Victor, P.; Cornelis, C.; De Derk, M. & Silva, P. P. (2009). Gradual trust and distrust in recommender systems. *Fuzzy Sets and Systems*, Vol. 160, No. 10, page numbers (1367-1382), ISSN: 0165-0114
- Wang, W. G.; Mokhta, M. & Linda, M. (2008). C-index: trust depth, trust breadth, and a collective trust measurement, *Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence*, pp. 13-16, ISBN: 978-1-60558-171-2, USA, June 2008, ACM Press, Pittsburgh, PA
- Yu, B.; Singh, M. P. & Sycara, K. (2004). Developing trust in large-scale peer-to-peer systems, *Proceedings of IEEE First Symposium on Multi-Agent Security and Survivability*, pp.1-10, ISBN: 0-7803-8799-6, USA, August 2004, IEEE Press, PA
- Yu, H. F.; Kaminsky, M.; Gibbons, P. B. & Flaxman, A. D. (2008). SybilGuard: Defending Against Sybil Attacks via Social Networks. *IEEE/ACM Transactions on networking*, Vol. 16, No. 3, page numbers (576-589), ISSN: 1-59593-308-5
- Zhang, Y. C. & Fang, Y. G. (2007). A Fine-Grained Reputation System for Reliable Service Selection in Peer-to-Peer Networks. *IEEE Transactions on Parallel and Distributed Systems*, Vol. 18, No. 8, page numbers (1134-1145), ISSN: 1045-9219
- Zhou R. F. & Hwang, K. (2007). PowerTrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing. *IEEE Transactions on Parallel and Distributed Systems*, Vol. 18, No. 4, page numbers (460-473), ISSN: 1045-9219

Wireless Technologies and Business Models for Municipal Wireless Networks

Zhe Yang and Abbas Mohammed
*Blekinge Institute of Technology
 Sweden*

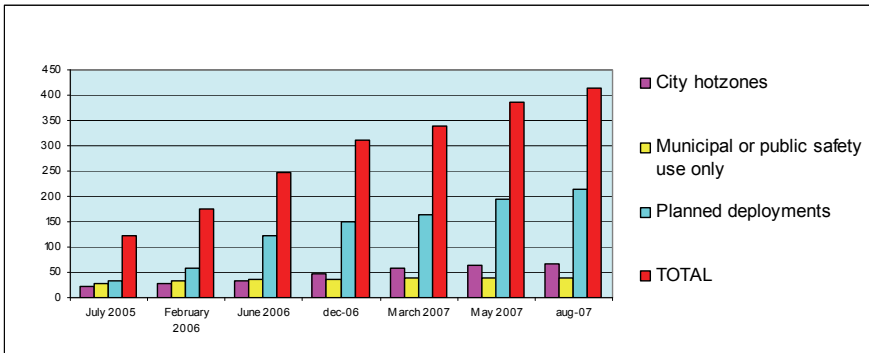
1. Introduction

Municipal wireless network is a recently established infrastructure of city-based wireless network, which provides mainly outdoor broadband wireless access to Internet for public usages. The municipal network is usually regarded as a public-utility service, which not only delivers well connected broadband services in the city at an affordable price but also promotes society interaction and brings sustainable development to support municipalities. Therefore, the concept is attracting more and more attention from city authorities both in developed and developing countries. Graph 1 shows the Internet usage worldwide in 2008 (Stats, 2008). It can be seen that there is a distinctive penetration rate between developed and developing regions.



Graph 1. Internet usage services worldwide by 2008

There are hundreds of cities, which have deployed and have plans to build municipal broadband networks over their territories. Graph 2 shows a tendency of wireless internet services covered by city-wide wireless network (Muniwireless, 2008).



Graph 2. An illustration of tendency to have wireless services in a city

City authorities are closely involved in network initiatives and rolling out with various forms and scales at different stages, because it is often argued that inexpensive or even free of charge broadband access network are impossible, or at least time-consuming to be realized by depending on market forces only. Generally, private network investors are cautious to protect investment, which could make the end goal of rolling out a full coverage area with an affordable price to be out of consideration. Therefore, employing a suitable business model of wireless city becomes an important choice regarding the basis and design of wireless city networks.

In this chapter, we first investigate and summarize existing and emerging concepts of business models for municipal networks implemented worldwide by distinguishing between ownership of the network infrastructure and service provisioning. The step demonstrates the way to find appropriate rationale and system architecture of municipal networks. To support our reasoning, we select examples of existing and emerging business models for municipal networks in different countries as our main focus to illustrate significant challenges to introduce municipal networks in a city and achieve low investments as well as open access to wireless services within affordable price for local residents.

2. Existing business models and examples of municipal network

2.1 Descriptions of business models for municipal networks

A proposed classification is constituted by all potential combinations between two key roles (network ownership and provisions) that can be taken up by common sorts of affiliation, such as public, private and a combination of them (MetroFi, 2008).

We follow the classification which fits the selected examples. In order to have a convincing reason, we consider both the ownership of network and service provisioning. Business roles are often taken up by different actors. Infrastructures of network operations are usually either associated with network ownership or service provisioning. Thus we can define following types of network in terms of the ownership of the infrastructure.

Private owner - the network is operated on the basis of a contractual arrangement in form of a license and concession. Therefore, municipality can deliver the rights of access to city's sites (streetlights, traffic lights, municipal buildings and so on), existing backbones (fiber, wired backhauls), as well as financial support.

Public owner - as the city authority that owns the network and operates it by using municipality enterprise funds to cover infrastructure costs.

Open site owner - the municipality provides open access to city's sites for the deployment of wireless network.

Next, we can define following types of networks in terms of service provisioning ownership.

Private owner - usually, a service provider, who supports and creates services to the network by gaining money from users' subscriptions and advertisements.

Public or Non-Profit owner - a provider, who allows an access to network services by using municipality's funding or applying for state or philanthropic grants.

Wholesale - can be consisted of a group of private owners, who offer and provide services to end users.

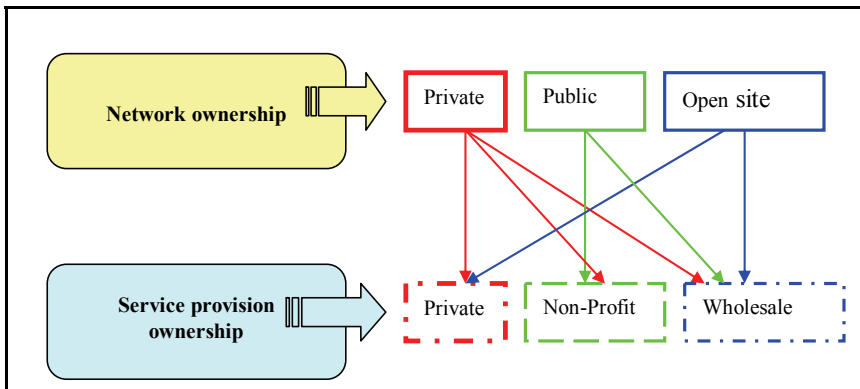


Fig. 1. Types of business model of municipal network providers (Yang et al., 2008)

Private-Private Model - In this model the key roles, such as network implementation, network operation as well as service providing, are performed by the same private company. The influence from the municipality can be limited in terms of financial supports or exclusive rent access to city assets. In return the private actor would provide limited Internet access to residents with low or no cost.

Private-Wholesale model - One of the most often implemented models in the US cities. The model is supported by deploying large mesh wireless clouds to deliver towards the public usage. Public entity's prospects of costs to deploy municipal networks are "nothing", though its profits are limited in rewarding from the network. Therefore, the public entity could retain a relative high level of influence, and take a less risk of claims to fairness among different service providers.

Private-Public or Non-Profit model - The model is mostly used for large mesh wireless clouds to offer services for public purposes. Financial inputs are from the authorities, therefore there is unlikely unfair service providers' competition.

Public-Public or Non-Profit model - The municipality builds and operates the network by itself in this model. All expenditures of network deployment and its operation are covered by the municipality, and service provision functions are also managed by the municipality. However, it is obvious that it is not an attractive model for the long term run.

Public-Wholesale model - The city builds and owns the municipal network. It signs an agreement with wholesale service providers, who can offer different contents with different

pricing subscriptions. This model is not widely implemented by large cities like Boston (Muniwireless, 2008).

Open site - Wholesale model - This model is similar to "Public-Wholesale" model. There is a difference that the access to build the network is granted either to the open sites in order to provide the wireless access to the community (city's inhabitants) or to limited region of the city for a particular group of users (low-income families, or only for the public safety purpose). The service provision is managed by a group of ISPs (Internet service providers) by allowing customers to access to the limited services (Metrofi, 2008).

2.2 Examples - Portland wireless city in USA

The business model of Portland wireless city can be interpreted as an "advertiser-supported" model (Muniwireless, 2008). The network is supported by the "private-wholesale" type of business relations, where network assets belong to MetroFi, the main network provider. Free web access services are supported by national and local advertisers. As an alternative, users who prefer an Internet access without advertisement can pay for premium services. Portland municipality becomes the major "anchor tenant" for MetroFi's wireless network provider. The main advantage of this public network is to allow all municipal employees to have an access to the network with certain functionalities. Fig. 2 gives an insight view of main actors' relations. Therefore, all functions in the public side belong to Portland's authority. However functions in the private part show only the influence to the outsourcing service providers, who provide advertising, consulting and customer-helpdesk support.

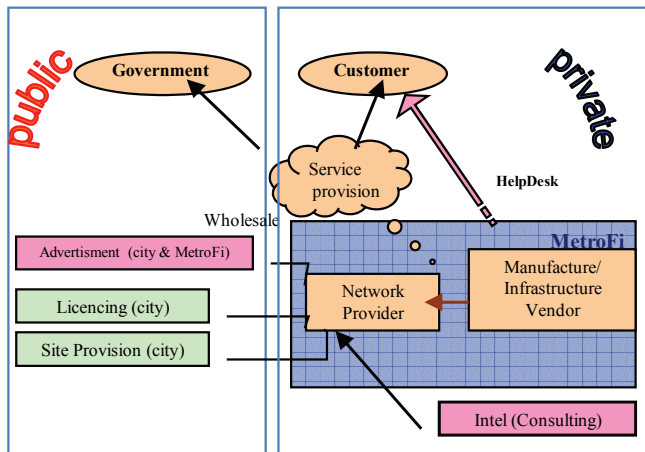


Fig. 2. Business model of Portland wireless City, USA

3. Emerging concepts of business model for Municipal networks

In middle 2007, Karlskrona, a city known as a historic naval and Telecom city on the southeast coast of Sweden, announced to have co-operations with *The Cloud*, the European wireless broadband network operator, to provide wireless city access in the city area, enhance an already rich suite of social services and improve innovations via services of the wireless city. A fixed fiber network operator, *Affärsverken Karlskrona AB* which is fully

owned by the Karlskrona commune, is on behalf of the municipality commune to cooperate with *The Cloud* to let the city be wireless (Affärsverken, 2006).

3.1 Wireless city of Karlskrona in Sweden and its motivations

The motivations and benefits of building a wireless city in Karlskrona have closely linked to the history and development strategies of the city. These are:

Making the city more attractive to IT and Telecom companies - A wireless enabled city can be more attractive to new IT and Telecoms companies, as well as facilitate business of companies in Karlskrona. Karlskrona, which was previously known as a 300-year-old fortress as well as an old ship-building yard, has now successfully created a new type of city based on its strategy to support IT and telecom industries after 1990s. Wireless city gives these companies a new approach to provide services as service providers. In addition, becoming a service partner of *The Cloud* can deliver their local services internationally without investing on the network infrastructures.

Delivery of social municipal and tourist services - Wireless city gives local residents more freedom to acquire information through wireless broadband at anywhere in anytime. It is easy to access public internet resources, such as transportation, education, leisure services, and society activities. Additionally, wireless city can assist tourists to access the local websites in Karlskrona via different Wi-Fi enabled terminals.

A natural expansion of the city's broadband network and increasing traffic - Wireless city can be naturally regarded as an expansion of the city fixed fiber network, which is owned and managed by *Affärsverken*. Since all the wireless network operators have to be the partner of *Affärsverken* in the business model, traffic passing through the fiber network is accordingly increases, therefore brings more revenue to this municipal company.

The main motivation for the wireless network operator *The Cloud* is the predicted increase usage of wireless network by local companies and residents. The sustainable strategy of city attracts growing attention and investment on IT and Telecom industries, and facilities internet connection for residents. For example, customers of *Telenor* can access *The Cloud's* network for free, and therefore generate traffic passing through the network of *The Cloud*. It is also predicted that mobile broadband is going to replace the fixed broadband in the future, which creates a profitable market for wireless network operator.

3.2 The business model and SWOT analysis

3.2.1 Implementation and strength

The business model implemented in the wireless city of Karlskrona can be generally categorized as the Public-Wholesale model (Cloud, 2007), where the local fixed network operator has established partnership with wireless network operator. Based on the partnership with *The Cloud*, *Affärsverken* can accordingly extend its fixed fibre network to include wireless infrastructure at nearly zero cost, and subsequently achieve the goal of wireless city. It has to be noticed that the Karlskrona municipality has branded the network as "*Wireless City of Karlskrona*" and fully owned the brand. It indicates that the municipality absolutely controls the wireless city network, and implements a neutral and open business model. New service providers and network operators can freely cooperate with *Affärsverken* and access to the business model.

In Karlskrona, *The Cloud* establishes the wireless network infrastructure and works actively with service providers, device providers and application partners to bring a range of services into its sites (Cloud, 2010). Consequently, *Affärsverken* can share the revenue of *The*

Cloud based on the traffic passing through its fiber network, and fully control activities of *The Cloud* for the purpose of a fair competition environment for different service providers and network operators. Fig. 3 shows the areas with wireless city service available in Karlskrona in 2007.

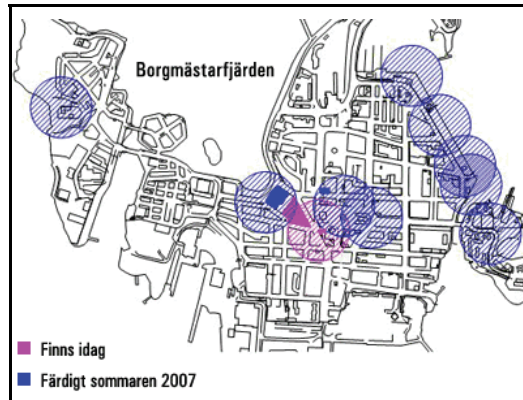


Fig. 3. Wireless city services available in the city centre of Karlskrona in 2007

Different roles of public and private actors involved in the business model are listed below:
Local authority - The Karlskrona municipality initializes the wireless city and provides funding to *Affärsverken* on behalf of the municipality to act as a public force to build wireless city. The local authority also gives access to buildings and light poles for mounting access equipment (Cisco, 2010).

Fixed network operator - *Affärsverken* provides backbone and shares revenue of the wireless network operator (Bar & Park, 2006). At the same time, it also acts as a regulator, which takes control over the network on behalf of the Municipality.

Wireless Network Operator - *The Cloud* mainly acts as a wireless network operator and brings services to end users. Being a network operator, it deploys and maintains the wireless network infrastructure in the wireless city. It is also responsible for managing and outsourcing the network capacity for service providers, and subsequently shares the revenue of service providers (Cisco, 2010). At the same time, *The Cloud* also provides an internet connection to end users paying through the Credit Card Company or mobile operator by sending messages through mobile phones.

Service Provider - It pays the network operator to let its customers to access the wireless network for free, or attracts new customers in the wireless city. In wireless city of Karlskrona, *The Cloud* has established partnership with various service partners, e.g. *Telenor*, *iPass*, *Spring PCS*, *Boingo*, *Trustive*, *Echovox SMS*, *AT&T*.

Technology Partner - It manufactures and sells devices to network operators as well as end users.

Credit Card Company or 3rd party - In wireless city of Karlskrona, it is responsible for charging and identifying end users, who don't have partner accounts of *The Cloud* in order to access the network. The payment module is widely used and gives a convenient way for users to subscribe services in the wireless city.

The business model of the wireless city in Karlskrona is shown in Fig. 4. By implementing the business model, Municipality can manage different actors in network rolling out, service

provision and revenue sharing. Based on the collaborations with *The Cloud*, Municipality has low investment on the wireless city infrastructure and low administrative burden. It fully achieves the goal of owning and controlling the network, open access for any wireless network operators and service providers, and delivering services at an affordable price for public access in the city.

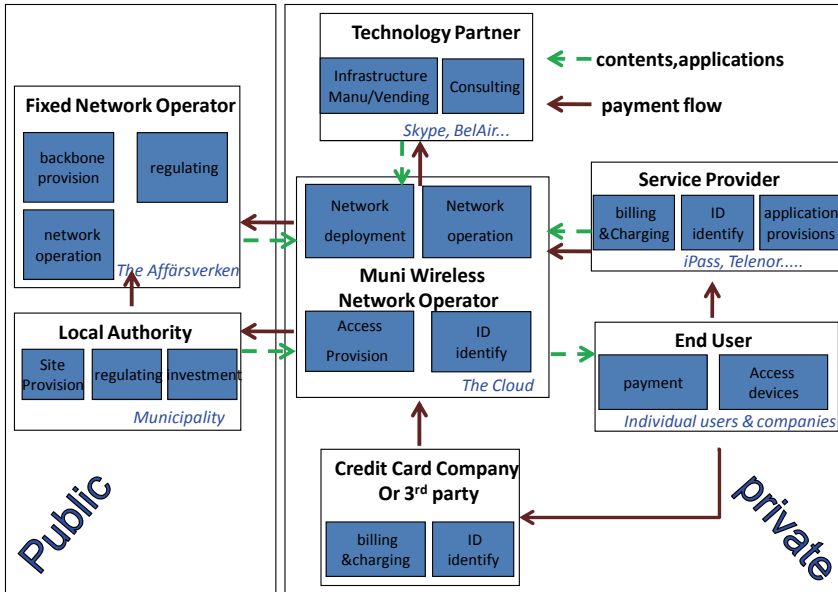


Fig. 4. Business model of wireless city in Karlskrona, Sweden

3.2.2 Weakness of the business model

The Cloud could not be a really open and fair network operator in the wireless city. Ideally, *The Cloud* should not be involved in selling internet services to end users as network operator. This distinct weakness is due to the natural conflict of *The Cloud* acting as an ISP and network operator. Being an ISP, *The Cloud* expects as many customers to buy service directly from it. However, as a network operator, *The Cloud* needs to sell access to other competitive ISPs. It can be hard for an ISP as well as a network operator not to discriminate against its competitors as well as customers. It is to some extent against the neutrality in the business model.

The Wireless City is collaborating with a potential competitor. Basically, there are no limitations to be accepted as a partner of the wireless city due to the revenue sharing module. Therefore, *The Cloud* can establish partnership with mobile network operator (MNOs) like *Telenor*, which can be regarded as the most potential competitors due to competitive 3G data technology and a large user subscriptions. These partners could terminate collaborations with *The Cloud*, and provide mobile broadband services as competitors.

The partnership may be not stable if some service providers are going to collaborate in the same type of business with new network operators. Currently *Skype* is a major service provider of *The Cloud* and accordingly is a service provider in the wireless city. However

customers now make *Skype* calls in some mobile handsets, which means it is possible to attract customers to use mobile broadband services from MNOs rather than wireless city.

3.3 Opportunities and threats driven by alternative wireless technologies

Cities can also be connected in a wireless fashion based on broadband infrastructures and services from fixed ISPs and mobile network operators, which are obvious competitors to Wi-Fi based municipal networks in terms of providing wireless connections in the cities areas.

Easy subscriptions, reasonable price with no binding offerings are provided by the wireless city of Karlskrona to give users more freedom to choose the time and period of subscribing services in order to attract more customers to access the network. Users have four alternatives to subscribe services from wireless city of Karlskrona:

- Paying to different service providers, e.g. *Telenor, iPass*.
- Buying airtime and making a payment through the Credit Card Company.
- Paying by sending a message via mobile operator network to acquire user name and password to log in.
- Buying coupons valid for one-day or seven-day at local coupon retailers in the city.

The price of accessing the municipal networks by paying through credit card for 30 minutes is SEK 40 and SEK 295 for 30 days. There is no binding and cancellation notice included in the offer. Compared with other ISPs in Karlskrona, e.g. the *Jacket Broadband AB, Tele2, Telia* and *Telenor*, the price offered by wireless city of Karlskrona is reasonable regarding to the broadband services available in most areas of the city. Users could also choose to be the customer of *The Cloud's* service partners, and therefore have free access to the network of wireless city. This module is suitable for local IT and Telecom companies to attract more customers by giving them free access in the wireless city as a reward.

Table 1 gives a comparison on subscriptions between wireless city and other fixed ISPs in Karlskrona. The information is collected from the websites of each operator in 2009.

Items/Provider	Wireless City	Jacket	Tele2	Telia	Telenor
Charge (Kr/Mon)	295	298	299	279	349
Speed limit (Mbps)	11 or 54	24	100	8	24
Connectivity	Wi-Fi	ADSL	LAN (ADSL)	ADSL	ADSL
Connection fee (Kr)	0	495	0	495	0
Binding period (Mon)	0	12	12	18	12
Cancellation notice in advance (Mon)	0	3	3	3	3
Mobility	Yes	No	No	No	No

Table 1. Subscription from fixed ISPs in Karlskrona

Since the wireless city service is not free, it is obvious to face a threat from mobile broadband services provided by MNOs. Table 2 shows a comparison on subscriptions between wireless city and MNOs in Karlskrona in terms of mobile broadband services. It can be seen that mobile broadband services from MNOs are more competitive in terms of mobility and coverage over the country. However, wireless city services can be more suitable for local residents and industries since speed and stable services are more important.

Items/Provider	Wireless City	3G	Tele2	Telia	Telenor
Charge (Kr/Mon)	295	199	189	229	199
Speed limit (Mbps)	11 or 54	7.2	7.2	7.2	24
Connectivity	Wi-Fi	3G	3G	3G WLAN	3G WLAN
Connection fee (Kr)	0	250	0	0	250
Binding period (Mon)	0	12	12	18	0~24
Cancellation notice in advance (Mon)	0	3	3	3	3
Mobility	Yes	Yes	Yes	Yes	Yes

Table 2. Subscriptions from MNOs in Karlskrona

Flexible and easy subscriptions, competitive pricing schemes and high-speed connection could make municipal networks in Karlskrona to be the most promising alternative to displace traditional fixed broadband and succeed ubiquitous mobility as a bonus.

4. Conclusions and future research

In this chapter, we have provided an overview of existing business models of wireless cities worldwide based on the ownership of network infrastructure and service provisioning. An example from Portland in US is given to illustrate an existing example of traditional advertisement-supported business model. Furthermore, we have introduced an emerging concept of business model and taken the wireless city of Karlskrona in Sweden as an example to illustrate main drivers, business actors, pricing and subscription schemes.

In general, the concept of wireless cities can not be treated as a pure business case since it has public and non-profit attributes. Based on our analysis, we come to the following conclusions:

- *Municipal initiative is essential.* Wireless city can be regarded as a symbol of a city and facilitate local activities. In our case, the Karlskrona municipality plays an important role to take the decision of building municipal networks based on local municipal profiles and development strategies.
- *Fair and open environment is more efficient for supporting competition among all parties involved.* Moreover, transparent business interactions inside municipal networks are mostly expected from the municipality. Whether being forced or volunteered to open its network, the municipal network operators need to provide opportunities for any ISPs and wireless network operators to be fairly associated into the network. It could be regarded as an emerging intention compared with traditional concept of wireless city, where a monopoly company occupies the most positions in the business model of the wireless city.
- *Wireless city services could be necessary to be free of charge.* Free services could be provided by municipal networks for people to acquire certain public information, e.g. public transportation timetable.
- *Low investments from a municipality could be achieved through the public-wholesale partnership business model.* In Karlskrona, the municipality can be regarded as a lossless actor in the market. It doesn't need to put much investment funds to the wireless network infrastructure, but it gains the privileges for its residents and local businesses.

The investment of all the actors involved in the business model is economical compared with the traditional monopoly model.

Based on our investigations, the emerging business model based on partnership collaboration is a suitable solution for regions to deploy municipal network. It could maximize users' choices, create a fair competition environment, remain the municipality as a leading regulator and activate all actors in the business model. In the future, it will be interesting to investigate the best combination of merging the emerging concepts into the traditional business scenarios. Below we list three interesting directions for further research.

- *Influence and regulation from municipality.* In the emerging concept, the municipality only forms a partnership with the network operator and highly involved in the network operation by delivering traffic through its fiber network. Furthermore, the municipality has a stronger influence on activities of municipal network operators. It will be interesting to explore the possibilities of introducing a second network operator as a competitor in two models to raise the competition or balance the influence from the existing operator.
- *A mixture of customer relationship with end users is also an interesting topic.* In traditional models, network operator acts more likely as an ISP and builds a direct relationship with end users. However the network operator in the emerging concept tries to avoid establishing a direct and long-term relationship with end-users in order to keep its neutral position for other ISP as its customers.
- *Types of collaboration with service providers involved in the model can be explored.* municipal network operator in the emerging concept has actively established extensive partnership with different service providers, but it acts actively as an ISP in the traditional business model. A combination of two types can be explored by forcing some service providers to access the municipal networks based on geographic divisions according to some agreements.

5. Reference

- Affärsverken AB. (2006). from www.affarsverken.se/privatkund/stadsnat/wireless-city
- Bar, F., & Park, N. (2006). Municipal Wi-Fi Networks: The Goals, Practices and Policy Implications of the US Cases. *Communications & Strategies*, vol 61, pp107-126, 2006
- Cisco (2010). Municipalities Adopt Successful Business Models for Outdoor Wireless Network, from <http://www.cisco.com>
- The Cloud. (2007). The Cloud Switches on Europe's most advanced WiFi Network across the City of London, from <http://www.thecloud.net/page/1796/About-us/Latest/Press-Releases/EN/The-Cloud-switches-on-Europe%3Fs-most-advanced-WiFi-network-across-the-City-of-London>
- The Cloud. (2010). Solutions for Locations, from <http://www.thecloud.net/For-business/Wireless-solutions/Solutions-for-locations/Cities>
- MetroFi (2008). from www.metrofi.com
- Muniwireless (2008). from www.muniwireless.com
- Internet World Stats. (2008). from www.internetworldstats.com
- Yang, Z., Khamit, S., Mohammed, A., & Larson, P. (2008). *A Comparative Study on Business Models of Municipal Wireless Cities in US and Sweden*. Paper presented at the *Third IEEE/IFIP International Workshop on Business-driven IT Management (BDIM)*.

Data-Processing and Optimization Methods for Localization-Tracking Systems

Giuseppe Destino, Davide Macagnano and Giuseppe Abreu
*University of Oulu - Centre for Wireless Communications
Finland*

1. Introduction

During the last years the increasing demand for location-based services (LBS) stimulated the developments of localization-tracking (LT) technologies. In this regard, a widely known and utilized LT technology is given by the Global Positioning System (GPS), that, based on satellite communications can achieve accuracy of tens of centimeters. However the inability to receive the satellite signal in indoor environments strongly limits the usage of GPS technology to outdoor scenarios. As a consequence, future indoor positioning services, *e.g.* surveillance, logistics and remote health-care applications, need to rely on the development of new LT systems based on short/medium-range wireless technologies. In this regard, thanks to their ability to provide accurate distance measurements under both LOS and non-LOS channel conditions, ultra-wideband (UWB) and chirp spread spectrum (CSS) technologies are two promising technologies to enable indoor LT services. In this chapter, we focus on technological and theoretical aspects of LT systems for indoor applications. In particular we consider two fundamental functionalities, namely, the data processing and the LT algorithm. We start describing an efficient wavelet-based filtering technique to process the data and to assess their reliability and we show how estimates on the confidence on the measurements can be used to improve the target locations. In the second part of the chapter we first cover some well known non-parametric state-of-the-art solutions to the localization problem, namely the classical-multidimensional scaling (MDS), the Nyström approximation and the SMACOF algorithm. Following we propose a novel low-complex technique that goes under the name of linear global distance continuation (L-GDC) and that we show to achieve the same performance of the maximum-likelihood estimator. Finally, the chapter ends with the simulation results and a short discussion on the open challenges.

2. Overview of a location-tracking system

In the near future it is expected that LT services will find usage in a very diversified range of scenarios. As an example, in the office environment illustrated in figure 1 it can be imagined that a LT system will be able to localize mobile or static terminals, *e.g.* notebooks, printers, PDAs and smart-phones using the radio access provided by an existing wireless network infrastructure (*e.g.* UWB, Wi-Fi, LTE-A, and so forth).

Assuming that all devices can communicate, up to their maximum radio range, amongst themselves and that a server is available to run the LT engine, the LT problem reduces to

find the position of the nodes in the network given a subset of the possible measurements amongst the devices.

To design a LT system, the first distinction that needs to be made is between the anchor nodes, whose location is assumed to be known to the system, and the target nodes, whose position needs to be estimated. For instance, referring to the office scenario depicted in figure 1, the N_A radio access points can be associated to the anchor nodes and the remaining N_T terminal devices with the targets.

At this point it is possible to introduce the three essential functional blocks characterizing of a standard LT system, namely, data acquisition, data processing and localization engine block illustrated in figure 2. The data acquisition block deals with the problem of extracting physical parameters such time-of-arrival (ToA), time-difference-of-arrival (TDoA), received-signal-strength (RSS) or angle-of-arrival (AoA) Mao et al. (2007) from the radios.

As shown in Li & Pahlavan (2004); Joon-Yong & Scholtz (2002), time-based metrics can provide distance measurements with few centimeters of error in line-of-sight (LOS) and tens of centimeters in non-LOS (NLOS) channel conditions.

Regarding AoA measurements, they are typically based on beam-forming or sensor array signal processing methods and they can achieve accuracies of few degrees in LOS conditions. In contrast to time-based mechanisms, AoA-enabled technology employs narrowband signal, and it is used to develop LT systems in large and open space environments.

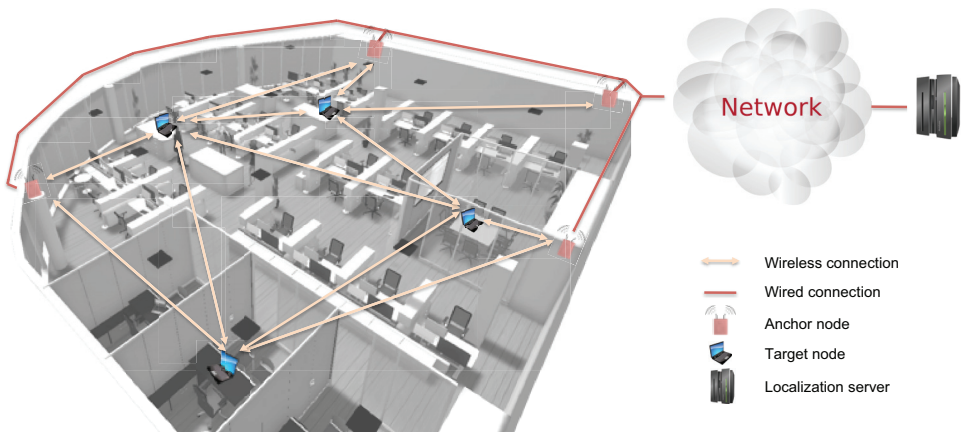


Fig. 1. Typical LT system in indoors

The third type of information is the RSS, which is a power-based metric and typically, available for both narrowband and wideband radio technologies. However the major problems with RSS measurements is their sensitivity to fading and the channel propagation model used. For these reasons, power-based LT systems are inaccurate, especially if RSS's are used to measure distances. Table 1 summarizes the features of the aforementioned metrics.

The second block in figure 2 deals with the processing of the observed data. In this block, a time-series filter is typically conceived to improve the signal to noise ratio of the measurements. The choice of the technique, however, depends on the dynamic and noise models assumed in the scenarios.

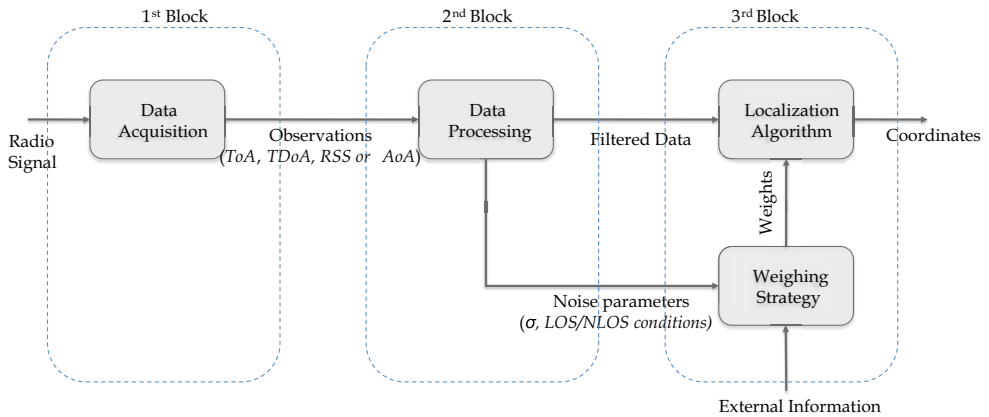


Fig. 2. Functional blocks of an LT system

For instance, in the case of a quasi-static target nodes and a stationary zero mean Gaussian noise affecting the observations even a simple moving average filter can substantially improve the location accuracy of the LT system.

Information Type	Strength	Weakness	Technology
ToA	High precision High multi-path resolution	Synchronization	Wideband
TDoA	High precision High multi-path resolution	Synchronized anchors Not suitable for mesh network topology	Wideband
RSS	Availability	Poor robustness to fading Model dependent	Wideband Narrowband
AoA	High resolution in LOS	Multiple antennas Sensitivity to NLOS	Narrowband

Table 1. Taxonomy of data acquisition methods

Finally, the third block concerns with the LT algorithm. In table 2, we provide an overview of the approaches typically utilized in the literature.

The first category is based on Bayesian formulations of the LT problem. The Kalman filter (KF) and its variations such as unscented, cubature and extended are, for instance, the most common Bayesian techniques utilized in the literature. They generally have good performance, although, they are very dependent on system and measurement models assumed in the problem formulation.

The second category of LT methods, instead, is based on non-parametric formulations of the problem. In contrast to Bayesian techniques, non-parametric approaches do not rely on any models and/or assumptions and because of this, the non-parametric methods are widely used in the literature Cox & Cox (2000); Costa et al. (2006); Destino & Abreu (2009); Shang & Ruml (2004); Cheung et al. (2004); Ouyang et al. (2010); Biswas, Liang, Toh, Wang & Ye (2006); Guvenc et al. (2008); Beck et al. (2008). This class of LT techniques can provide very accurate location estimates but they often involve the optimization of a non-convex objective

function. The last category of LT techniques deals with fingerprint methods. The fundamental idea is to search the best pattern match between the stored data and the observations. Fingerprint methods are two-phases approaches. The first one is to construct a database using a priori measurements of the considered parameter (RSS, power profile, ToA, etc.) at different locations, while the second one is to search for the best pattern match between observations and data. This type of LT-technique is very practical in many application scenarios, adapts well with both time- and power-based metrics, but it is sensitive to changes of the environment.

LT Algorithm	Description	Assumption	Weakness
Bayesian	Estimation based on the <i>a posteriori</i> pdf	System and measurement models	Dependency on the reliability of the model
Non-parametric	Optimization of on the least-squared error function	None	Typically difficult optimization problem
Fingerprint	Two-phase approach	Map of the physical parameters in the coverage area	Dependency on the accuracy of the map

Table 2. Classification of LT algorithms.

3. Distance-based non-parametric LT system

From now on we focus on distance-based non-parametric LT approaches, where the distance measurements (ranging) are assumed to be the output of either ToA or RSS measurements.

After introducing the basic formulation of the localization problem, we will describe a wavelet based filter to smooth the raw-observations and state-of-the-art optimization methods to minimize the least-squared objective function used in the formulation of the problem.

3.1 LT problem statement

Consider a network deployed in the η -dimensional space and let $\mathbf{X} \in \mathbb{R}^{N \times \eta}$ denote the coordinate matrix, whose i -th row-vector $\mathbf{x}_i \in \mathbb{R}^\eta$ indicates the location of the i -th node. The set of indexes $\{i \leq N_A\}$ and $\{N_A < i \leq N\}$ refer to anchors and targets, respectively. Let \mathbf{D} indicate the Euclidean distance matrix (EDM) of \mathbf{X} obtained from the Euclidean distance function (EDF) $\mathcal{D}(\mathbf{X}) : \mathbb{R}^{N \times \eta} \rightarrow \mathbb{R}^{N \times N}$ defined as follows Dattorro (2005)

$$\mathbf{D} = \mathcal{D}(\mathbf{X}) \triangleq \sqrt{\mathbf{1}_N \cdot \text{diag}(\mathbf{X}\mathbf{X}^T)^T + \text{diag}(\mathbf{X}\mathbf{X}^T)\mathbf{1}_N - 2\mathbf{X}\mathbf{X}^T}, \quad (1)$$

where $\mathbf{1}_N$ is a column vector of N elements equal to 1, and $\text{diag}(\cdot)$ indicates a column vector containing the diagonal elements of its argument.

The (i, j) -th element of \mathbf{D} , denoted by d_{ij} , is the Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|_F$ between the pair of nodes $(\mathbf{x}_i, \mathbf{x}_j)$, where $\|\cdot\|_F$ indicates the Frobenius norm.

The k -th EDM sample $\tilde{\mathbf{D}}_k$ of the set of EDM samples $\{\tilde{\mathbf{D}}_k\}$, are composed of measurements of d_{ij} , denoted by $\tilde{d}_{ij,k}$ described by the ranging model

$$\tilde{d}_{ij,k} = d_{ij} + n_{ij,k}, \quad (2)$$

where $n_{ij,k}$ is assumed to be a Gaussian random variable with zero mean and variance σ_{ij}^2 . Let $\mathcal{M}: \mathbb{R}^{N \times N \times K} \rightarrow \mathbb{R}^{N \times N}$ denote the functional model of the data-processing block and let $\bar{\mathbf{D}}$ be the smoothed EDM computed as

$$\bar{\mathbf{D}} \triangleq \mathcal{M}(\{\tilde{\mathbf{D}}_k\}), \quad (3)$$

where K is the total number of EDM samples.

Then a non-parametric formulation of the LT problem is given by the following weighted least square (WLS) minimization

$$\begin{aligned} \hat{\mathbf{X}} = \arg \min_{\mathbf{X} \in \mathbb{R}^{N \times \eta}} \quad & \frac{1}{2} \cdot \left\| \mathbf{W}^\circ \left(\bar{\mathbf{D}}^{\circ q} - \mathcal{D}(\hat{\mathbf{X}})^{\circ q} \right) \right\|_{\mathbb{F}}^2, \\ \text{subject to} \quad & \hat{\mathbf{x}}_i = \mathbf{x}_i \quad \forall i = 1, \dots, N_A \end{aligned} \quad (4)$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{N \times \eta}$ indicates an estimate of \mathbf{X} , \mathbf{W} is a weighing matrix that relates to the reliability of $\bar{\mathbf{D}}$, q is an exponent typically chosen amongst the values $\{1, 2\}$ and \circ indicates the point-wise (Hadamard) power or product, respectively.

The proposed WLS approach is widely used in the literature for several reasons. First, under the assumption of $N_T = 1$ and $q = 2$, the exact solution of the minimization problem can be computed with a close-form algorithm Beck et al. (2008). Second, under the assumption of low noise, the WLS objective function can be linearized without compromising the accuracy of the location estimates Cheung et al. (2004); Guvenç et al. (2008). Third, under the assumption of a zero-mean Gaussian noise and $q = 1$, the WLS approach is equivalent to the maximum-likelihood (ML) formulation of the LT problem Patwari et al. (2003); Biswas, Liang, Toh & Wang (2006). Indeed, by computing the likelihood function of $\hat{\mathbf{X}}$

$$p(\hat{\mathbf{X}} | \bar{\mathbf{D}}) = \frac{1}{(2\pi)^\eta} \prod_{e_{ij} \in E} \exp \left(-\frac{(\bar{d}_{ij} - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_2)^2}{\sigma_{ij}^2 / K_{ij}} \right), \quad (5)$$

where K_{ij} is the number of measurements of d_{ij} , e_{ij} indicates an connected link between the i th and the j -th nodes, and E as the set of all connected link, and taking the logarithm it follows that

$$\ln(p(\hat{\mathbf{X}} | \bar{\mathbf{D}})) = \frac{1}{(2\pi)^\eta} \sum_{e_{ij} \in E} \frac{K_{ij}}{\sigma_{ij}^2} \left(\bar{d}_{ij} - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_2 \right)^2, \quad (6)$$

which is equivalent to equation 4 rewritten as

$$\sum_{e_{ij} \in E} w_{ij}^2 \cdot \left(\bar{d}_{ij}^q - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_2^q \right)^2, \quad (7)$$

with $w_{ij}^2 = \frac{K_{ij}}{\sigma_{ij}^2}$ and $q = 1$.

4. Ranging post-processing

As shown in figure 2, while block-1 deals with the detection, acquisition and association problem, block-2 pre-filters the data to improve the signal to noise ratio for the measurements. Although several algorithms can be used to smooth the observations, e.g. moving average, exponential, autoregressive moving average, Kalman filters and so forth, here after we will focus on a low-complex wavelet-based pre-filtering that has been proved to suit the localization and target tracking scenarios.

4.1 Wavelet-based smoothing

The Wavelet Transform (WT) makes use of a unique dilated window (the wavelet function) to analyze signals. This allows good time resolution (for short windows) at high frequency and good frequency resolution (corresponding to long-window) at low frequency S.Mallat (1998). The decomposition is based on a family of functions $\sqrt{s}\psi(s(x-u))_{(s,u)\in\mathbb{R}^2}$ corresponding to the translated and dilated version of the *wavelet* function $\psi(x)$, also called *mother wavelet*, and with s and u corresponding to the *scaling* and *translation* factors. Given a continuous function $f(x)$, its continuous wavelet transform, here denoted as $Wf(s,u)$, corresponds to the inner product $\langle f(x), \psi_s(x-u) \rangle$, meaning the cross-correlation between the original function and the scaled wavelet shifted at u . In S.Mallat & S.Zhong (1992) it is shown that choosing $\psi(x) = d\phi(x)/dx$, with $\phi(x)$ as smoothing function, then it is possible to characterize the shape of irregular functions $f(x)$ through $Wf(s,u)$. In addition, using the properties of the convolution operator it follows that

$$Wf_s(x) = f * \left(s \frac{d\phi_s}{dx} \right) (x) = s \frac{d}{dx} (f * \phi_s)(x), \quad (8)$$

which allows to interpret $Wf(s,u)$ as the derivative of a local average of $f(x)$, with smoothing degree depending on the scale factor s .

Amongst the several algorithm to compute the wavelet transform of discrete signals, because of its low complexity and its redundant representation of the signal $f(x)$ across the scales, which has been proved to be particularly suitable in filtering applications, we use the *à trous* algorithm briefly summarized in the following.

Let $a_0[n]$ be the discrete signal to be analyzed, with n as the discrete time index and assume the value for $a_0[n]$ in n equivalent to the local average between the original continuous function $f(x)$ and a kernel function $\phi(x-n)$ (namely $\langle f(x), \phi(x-n) \rangle$), then at any scale $j > 0$, a smoothed version of $a_0[n]$ is computed as $\langle f(x), \phi_{2^j}(x-n) \rangle$, with

$$\phi_{2^j}(t) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{x}{2^j}\right). \quad (9)$$

The function $\phi(x)$ is called *scaling function* and it corresponds to a low-pass filter, while the coefficient for the dyadic WT are obtains by $z_{2^j}[n] = Wf(s,n) = \langle f(x), \psi_{2^j}(x-n) \rangle$ with $\psi_{2^j}(x-n)$ defined similarly to equation 9. Once the low-pass filter $h[n]$ and high-pass filter $g[n]$ are designed then $a_0[n]$ is decomposed by repetitively computing

$$\begin{aligned} a_{j+1}[n] &= a_j * \tilde{h}_j[n], \\ d_{j+1}[n] &= a_j * \tilde{g}_j[n], \end{aligned} \quad (10)$$

with $h_j[n]$ obtained from $h[n]$ inserting $2^j - 1$ zeros between each sample of the filter (similarly for $g_j[n]$) and $0 < j < J$.

We use the WT to study each time series corresponding to subsequent ranging measured at the devices. We restrict ourselves to LoS target tracking scenarios, and we suggest a scheme to adaptively pre-filtering the observations $f[n]$ in a completely non-parametric fashion. To do so we use the output of the DWT to estimate σ_d and δ , namely the noise level affecting the observations and the target dynamic perceived at each anchor via ranging.

As mentioned above, the wavelet coefficients $d_j[n]$ computed at the different scales j include the high frequency components for the original signal and it is therefore used to characterize σ_d . Similarly, the output of the scaling function is used to infer δ . Clearly the approach works at best if the signal can be decomposed in high frequency components of short duration and a low frequency part of relatively long duration. From equations 8 and 10 it is clear that the wavelet coefficients $d_j[n]$ at the scale $j = 1$ represents a simple differential operator, therefore under static ($\hat{v} = 0$) or anyway scenario characterized by a small dynamic \hat{v} , an estimate of $\hat{\sigma}_d$ can be inferred from $d_1[n]$.

However, when the target dynamic (\hat{v}) increases, $d_1[n]$ starts including part of the energy associated to $f[n]$ and eventual estimates of $\hat{\sigma}_d$ would be affected by error. To overcome this problem $\hat{\sigma}_d$ is estimated as the standard deviation of $d_1[n]$ computed from subsets of subsequent observations characterized by the same polarity value in the support function

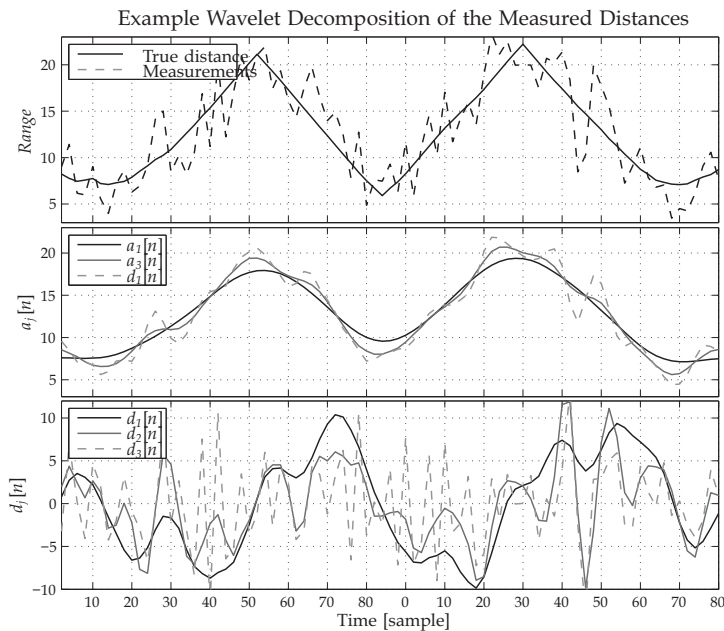


Fig. 3. Example wavelet decomposition of TOA ranging.

described in section 5.1.2. To distinguish the long term time process associated to the real TOA measurements (low-pass filtered version of $f[n]$), we use an averaged version of $a_1[n]$ in the same way proposed in Macagnano & de Abreu (2008), meaning that at each sampling time we compute the DWT on a window of size 2^l and centered at n .

From this averaged $a_1[n]$ we compute the parameter δ approximating the perceived dynamic at the specific anchor with respect to the considered target. The decomposition of $f[n]$ in its high/low frequency components is performed subject to the boolean operator Θ defined in section 5.1.2.

Using Θ , computed at each time n and for each anchor-to-target link, we decide whether the real ToA observation is better approximated by the measured ranging ($f[n]$) or its low-pass filtered version ($a_1[n]$). The only price paid using this wavelet smoothing based on Θ , is the introduction of a lag of $2^l - 1$ samples in the computation.

5. LT algorithm

The LT problem formulation considered in this chapter is the WLS-ML approach

$$\begin{aligned} \hat{\mathbf{X}} = \arg \min_{\hat{\mathbf{X}} \in \mathbb{R}^{N \times \eta}} \sum_{e_{ij} \in E} w_{ij}^2 \cdot (\bar{d}_{ij} - \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_{\mathbb{F}})^2 \\ \text{subject to } \hat{\mathbf{x}}_i = \mathbf{x}_i \quad \forall i = 1, \dots, N_A. \end{aligned} \quad (11)$$

The challenges faced in this optimization problem are: the computation of the weights and the minimization of the objective function. In the sequel, we will tackle both issues and we will describe in details very effective solutions.

5.1 Weighing strategies

In the optimization problem posed in equation 11, the purpose of the weights is "to reflect differing levels of concern about the sizes of the terms" in the objective function. In other words, higher the weight tighter is the concern Boyd & Vandenberghe (2004).

The first proposed strategy, nevertheless optimal in the ML sense, can be derived directly from equation 6,

$$w_{ij}^* = \frac{K_{ij}}{\sigma_{ij}^2}, \quad \forall \sigma_{ij}^2 \neq 0. \quad (12)$$

For the special case of $\sigma_{ij}^2 = 0$, i.e. no error, $w_{ij}^* = 0$ but in equation 11, we add the equality constraint

$$\hat{d}_{ij} = \tilde{d}_{ij}. \quad (13)$$

In most cases, however, σ_{ij}^2 's are not known *a priori*, therefore, alternative weighing strategies will be considered. The first alternative referred to as binary weight or unweighted is

$$w_{ij}^u = \begin{cases} K_{ij}, & \forall e_{ij} \in E, \\ 0, & \text{otherwise} \end{cases}. \quad (14)$$

This method is also optimal if $\sigma_{ij}^2 = \sigma^2 \forall ij$ (ij because $\frac{1}{\sigma}$ becomes a constant and therefore, it is a common factor to the WLS-ML objective function).

The second alternative is to replace in equation 12 σ_{ij}^2 by the sample variance $\hat{\sigma}_{ij}^2$ estimated either in the wavelet filter as described in section 5.1.2 or computed as

$$\hat{\sigma}_{ij}^2 = \frac{1}{K_{ij} - 1} \sum_{k=1}^{K_{ij}} \left(\tilde{d}_{ij,k} - \bar{d}_{ij} \right)^2, \quad (15)$$

where \bar{d}_{ij} is the sample mean computed as

$$\bar{d}_{ij} = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \tilde{d}_{ij,k}. \quad (16)$$

A third method is based on the regression model proposed in Costa et al. (2006). This technique is very effective only if a sufficient number of measurements are collected and if short distances are more reliable than long ones. In this case, weights are given by

$$w_{ij}^e = \sum_{k=1}^{K_{ij}} \exp \left(\frac{-\tilde{d}_{ij,k}^2}{\left(\max_j \tilde{d}_{ij,k} \right)^2} \right), \forall e_{ij} \in E. \quad (17)$$

Yet another alternative weighing strategy, which will be described with more details in the following subsection, is based on the relationship between the concept of concern (seen as constraint in the optimization) and the notion of statistical confidence Destino & De Abreu (2009). In essence, the weight, hereafter referred to as dispersion weight, will be computed as "a measure of the confidence about the estimate \bar{d}_{ij} associated with the penalty on the assumption of LOS conditions".

5.1.1 Dispersion weights

The dispersion weight is mathematically formulated as

$$w_{ij}^D \triangleq \Pr \left\{ -\gamma \leq \varepsilon_{ij} \leq \gamma \right\} \cdot \mathcal{P}_{ij}, \quad (18)$$

where γ is the confidence bound of the observation \bar{d}_{ij} around the true distance d_{ij} , \mathcal{P}_{ij} is a penalty imposed over the LOS assumption and $\varepsilon_{ij} \triangleq d_{ij} - \bar{d}_{ij}$.

For convenience, we shall hereafter use w_{ij}^L in reference to the probability in equation 18, such that $w_{ij} = w_{ij}^L \cdot \mathcal{P}$. The weights w_{ij}^L and \mathcal{P}_{ij} will also be dubbed the confidence weights and penalty weights, respectively.

Under the assumptions that $\rho_{ij} = 0$, i.e. LOS channel conditions, and $\tilde{d}_{ij,k}$ are independent, the dispersion weight can be rewritten in the form

$$w_{ij}^L = 2 \cdot \Pr \left\{ \varepsilon_{ij} \leq \gamma \right\} - 1, \quad (19)$$

where we use the LOS assumption to set $\mathcal{P}_{ij} = 1$.

Considering $\tilde{d}_{ij,k}$ as Gaussian random variable, by means of small-scale statistics w_{ij}^L can be computed as Gibbons (1992),

$$w_{ij}^L(\hat{\sigma}_{ij}, K_{ij}; \gamma) = -1 + 2 \cdot \int_{-\infty}^{T_{ij}} f_T(t; K_{ij} - 1) dt, \tag{20}$$

$$T_{ij} = \gamma \cdot \sqrt{K_{ij} / \hat{\sigma}_{ij}^2}, \tag{21}$$

where $f_T(t; n)$ is the T -distribution of n degree of freedom and T_{ij} is the t -score. As emphasized by the notation, w_{ij}^L is a function of the sample variance $\hat{\sigma}_{ij}^2$ and the number of samples K_{ij} , as well as the confidence bound γ , to be specified below. Since $\hat{\sigma}_{ij}^2$ and K_{ij} carry different information about the true value of d_{ij} , it is not surprising that both these parameters impact on the weight w_{ij}^L . In fact, the plots of $w_{ij}^L(\hat{\sigma}_{ij}, K_{ij}; \gamma)$ illustrated in figures 4 and 5, show that w_{ij}^L grows with the inverse of $\hat{\sigma}_{ij}^2$ (for fixed K_{ij}), and with K_{ij} (for fixed $\hat{\sigma}_{ij}^2$). This is in accordance with the argument outlined in the heading of this section and widely invoked by other authors Biswas, Liang, Toh, Wang & Ye (2006); Costa et al. (2006); Shang & Ruml (2004); Patwari et al. (2003); Boyd & Vandenberghe (2004); Alfakih et al. (1999), since $\hat{\sigma}_{ij}^2$ is proportional to the uncertainty of \tilde{d}_{ij} , as a measure of d_{ij} , while K_{ij} relates to the quality of \tilde{d}_{ij} and $\hat{\sigma}_{ij}^2$ as measures of d_{ij} and its dispersion, respectively. Unlike $\hat{\sigma}_{ij}^2$ and K_{ij} , which are obtained in the process of measuring inter-node distances, the confidence bound γ is a free choice parameter that allows for fine-tuning the relative values of w_{ij}^L .

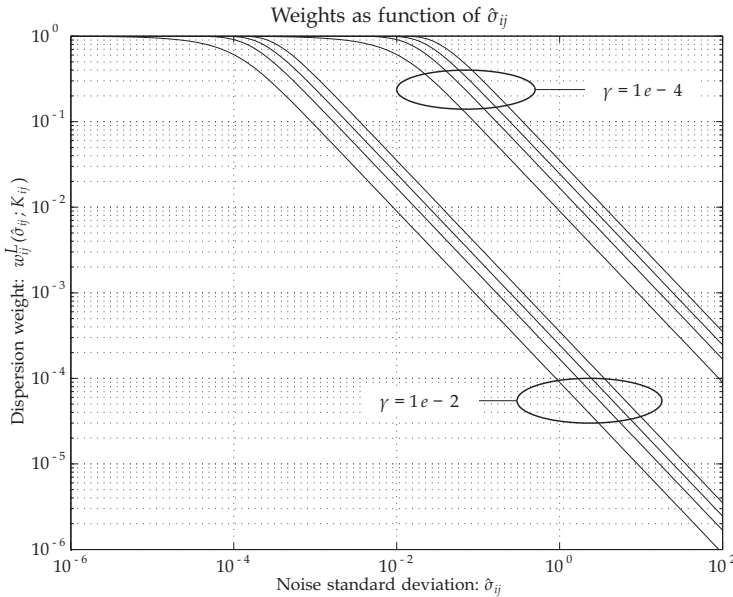


Fig. 4. Dispersion weight as functions of $\hat{\sigma}_{ij}$

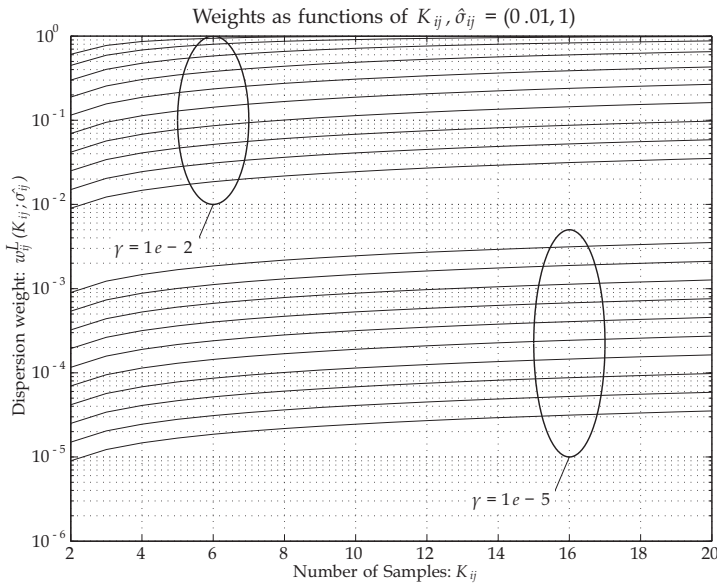


Fig. 5. Dispersion weight as functions of K_{ij}

The mechanism to find the optimum γ is given by the following optimization problem based on the *diversity index* or *entropy* metric

$$\gamma_{\text{opt}} = \arg \max_{\gamma \in \mathbb{R}^+} \mathcal{H}(\gamma), \tag{22}$$

where

$$\mathcal{H}(\gamma) = - \sum_{k=K_{\min}}^{K_{\max}} \int_{\sigma_{\min}}^{\sigma_{\max}} w^L(s, k; \gamma) \cdot \ln(w^L(s, k; \gamma)) ds, \tag{23}$$

where K_{\min} , K_{\max} , σ_{\min} and σ_{\max} are the minimum and the maximum number of observable samples and the minimum and the maximum typical ranging error, respectively.

The derivation of the method is omitted in this book, however, an interested reader can refer to Destino & De Abreu (2009). To validate the aforementioned optimization criterion, in figure 6 we show that varying γ , the minimum root-mean-square-error obtained via solving the optimization in equation 11 is close to that one achieved with γ_{opt} .

For the sake of completeness, in figure 7 we illustrates γ_{opt} as function of K_{\max} and σ_{\max} , considering $K_{\min} = 2$, $\sigma_{\min} = 1e-4 \approx 0$. The same results are also shown in table 3.

5.1.2 Dynamic weighing strategy

In this section it is shown how to use the output of the wavelet transform of the time series $f[n]$ corresponding to the ToA measurements at each anchor node to extract a confidence on the observations in the form of

$$w_i[n] = \frac{1}{\hat{\sigma}_d}. \tag{24}$$

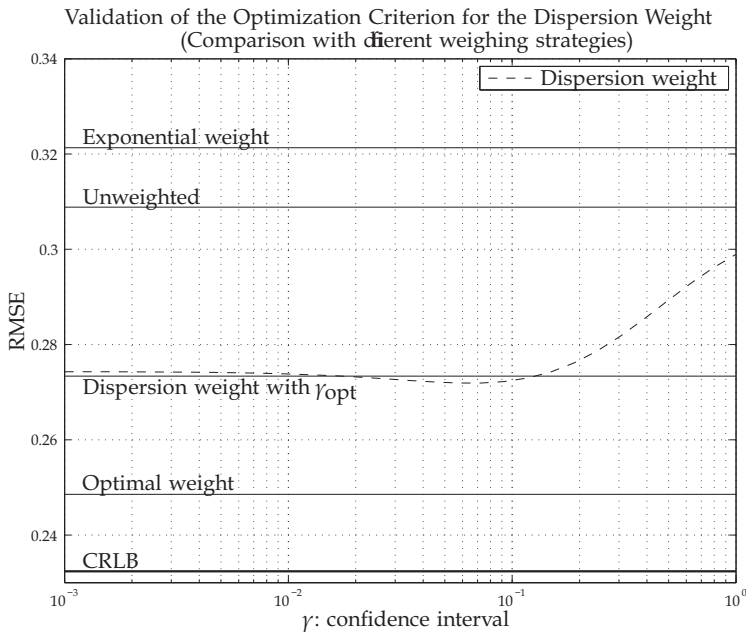


Fig. 6. Validation of the optimization criterion of γ . Simulation results are obtained for a network with $N_A = 4$, $N_T = 1$, $\sigma_{min} = 1e-4$, $\sigma_{max} = 2$, $K_{min} = 2$ and $K_{max} = 7$.

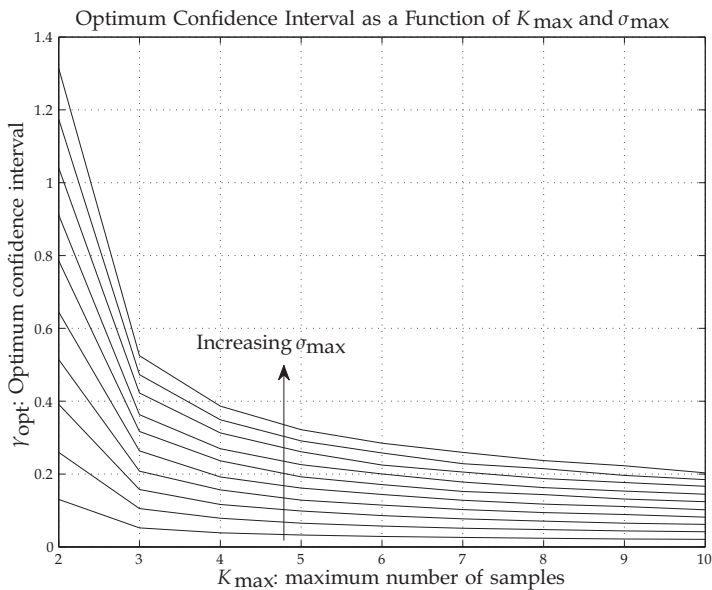


Fig. 7. Plot of the optimal confidence interval as a function of σ_{max} and K_{max} , with $\sigma_{min} = 1e-4$ and $K_{min} = 2$.

K_{\max}	σ_{\max}									
	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	1.80	2.00
2	13.04	25.94	39.13	51.41	64.49	78.58	91.10	104.04	117.45	131.39
3	5.20	10.51	15.75	20.79	26.36	31.67	36.28	42.25	47.28	52.48
4	3.86	7.85	11.63	15.66	19.19	23.62	26.96	31.29	34.91	38.65
5	3.24	6.52	9.85	12.85	16.14	19.24	22.56	26.13	29.10	32.16
6	2.86	5.70	8.58	11.47	14.38	17.11	20.03	22.49	25.78	28.46
7	2.59	5.13	7.69	10.24	12.81	15.22	17.79	20.55	22.83	25.96
8	2.35	4.74	7.08	9.41	11.75	14.35	16.27	18.78	21.49	23.69
9	2.18	4.37	6.51	8.89	11.09	13.14	15.33	17.68	19.62	22.28
10	2.07	4.15	6.16	8.17	10.17	12.39	14.45	16.65	18.46	20.33

Table 3. Tabulation of the optimal confidence interval γ_{opt} . Values are indicated with the multiplicative factor $1e-2$.

Obviously the problem is to recover a sufficiently good estimate of $\hat{\sigma}_d$ from the time varying process $f[n]$. To do so we use a boolean support to distinguish, amongst subsequent observations of $f[n]$, the one belonging to the same trend. To do it, we check whether the coefficient $d_j[n]$ at n and computed as accordingly to subsection 4.1, retain the same polarity across the scales $2 < j < J$. Whether the condition above is satisfied, then we assign to the n th value of our support ± 1 , accordingly to the polarity of $d_1[n]$, differently we assign 0. The reason to distinguish the oscillatory behavior of $f[n]$ using the multiresolution representation provided by $d_j[n]$ ($2 < j < J$) is that it better characterizes the trend of $f[n]$, since it allows to assign 0 to the points corresponding to significantly high discontinuity on $f[n]$, such as direction changes in the targets trajectories. Thus, computing the support above at each time n and of each anchor-to-target link, it is possible to distinguish amongst subsequent measurements, the one belonging to the same set of subsequent points in the support, and therefore mainly affected by the energy of the noise process.

Consequently $\hat{\sigma}_d$ is estimated as the standard deviation for $d_1[n]$, from those subsets of subsequent observations characterized by the same polarity value in the support. In case of the support equal to 0 at n , then we simply keep the previous estimate. However, since $d_1[t]$ is computed convolving $f[t]$ with $g_1[n]$, then the estimated $\hat{\sigma}_d$ described in section 4.1 needs to be compensated accordingly to the values of $g[n]$, which by $\hat{\sigma}_j^e = \hat{\sigma}_d \sigma_j$, with $\hat{\sigma}_j^e$ as the value estimated directly from $d_j[t]$ and σ_j a reference value pre-computed and dependent on the wavelet used, namely the filter $g[n]$. Therefore $\hat{\sigma}_d$ is computed from $d_1[n]$ from the points belonging to the same subsets on the support as

$$\hat{\sigma}_d = \hat{\sigma}_j^e / \sigma_j. \quad (25)$$

A limit on the minimum number points representing the subsets has to be fixed. In our simulation this value was fixed to 3. To control whether $f[n]$ can be decomposed in high frequency components of short duration and a low frequency part of relatively long duration, meaning that we test whether it is feasible to approximate $f[n]$ using its low-pass filtered version ($a_1[n]$), we use the boolean operator Θ defined by

$$\begin{cases} \delta > \sigma_d \rightarrow \Theta = 1, \\ \delta < \sigma_d \rightarrow \Theta = 0. \end{cases} \quad (26)$$

5.2 Optimization methods

After the calculation of the weights, the next challenge is to minimize the objective function in equation 11. In the sequel, we will describe state-of-the-art as well as novel optimization methods to compute reliable solutions.

5.2.1 Classical-multidimensional scaling

The classical-multidimensional scaling (CMDS) approach Cox & Cox (2000) can be thought of as an algebraic solution of the ML-WLS localization problem with $\mathbf{W}=\mathbf{1}_N \cdot \mathbf{1}_N^T - \mathbf{I}_N$, where $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ indicates the identity matrix. Under the assumption that all d_{ij} are measured at least once ($K_{ij} \neq 0 \forall (i, j)$), this solution is the least squares solution of equation 4 Cox & Cox (2000). The fact that no iterations are required makes the CMDS a very low-complexity and fast solution of the localization problem. It can be shown, however, that in the presence of incomplete EDM samples, the performance of the CMDS algorithm degrades drastically. Therefore, in the context of incomplete mesh network topology, the CMDS or some its variations, will be used to initialize iterative minimization techniques.

Briefly, the CMDS algorithm can be summarized as follows. First, compute the kernel $\bar{\mathbf{K}}$

$$\bar{\mathbf{K}} \triangleq \mathcal{K}(\bar{\mathbf{D}}) = -(\mathbf{T} \cdot (\bar{\mathbf{D}})^{\circ 2} \cdot \mathbf{T}) / 2, \quad (27a)$$

$$\mathbf{T} \triangleq \mathbf{I}_N - (\mathbf{1}_N \cdot \mathbf{1}_N^T) / N. \quad (27b)$$

Then, an estimate of the node coordinates $\hat{\mathbf{X}}$ is obtained as

$$\hat{\mathbf{X}} = \left([\mathbf{U}]_{\text{UL}; N \times \eta} \cdot [(\Lambda)^{\frac{1}{2}}]_{\text{UL}; \eta \times \eta} \right)^T, \quad (28)$$

where $[\cdot]_{\text{UL}; n \times q}$ denotes the n -by- q upper-left partition and the matrices \mathbf{U} and Λ are the eigenvector and eigenvalue matrices of $\bar{\mathbf{K}}$ Cox & Cox (2000), respectively, both in decreasing order.

5.2.2 Nyström algorithm

An alternative to the CMDS is the *Nyström approximation*¹ technique Williams & M.Seeger (2000); C. Fowlkes & Malik (2004). This method performs the same eigen-decomposition of CMDS, but in a more efficient manner.

Consider the Nyström kernel given by C. Fowlkes & Malik (2004)

$$\tilde{\mathbf{K}} \approx \left[\begin{array}{c|c} [\mathbf{K}]_{1:\eta, 1:\eta} & [\mathbf{K}]_{1:\eta, \eta+1:N} \\ \hline [\mathbf{K}]_{1:\eta, \eta+1:N}^T & [\mathbf{K}]_{1:\eta, \eta+1:N}^T \cdot [\mathbf{K}]_{1:\eta, 1:\eta}^{-1} \cdot [\mathbf{K}]_{1:\eta, \eta+1:N} \end{array} \right], \quad (29)$$

in which $[\mathbf{K}]_{1:\eta, 1:\eta}$ and $[\mathbf{K}]_{1:\eta, \eta+1:N}$ denote the upper-left η -by- η , and the upper-right η -by- $(N - \eta)$ minors of \mathbf{K} .

Recall that (Dattorro, 2005, pp. 195)

¹ In the case of Euclidean kernels, the Nyström “approximation” is actually an exact completion if the entries of the required minors are error-free.

$$[\mathbf{K}]_{1:\eta,1:\eta} = -\frac{1}{2} \left([\mathbf{D}]_{1:\eta,1:\eta} + \mathbf{C}_1 \otimes \mathbf{1}_\eta \mathbf{1}_\eta^T - \mathbf{C}_2 \otimes \mathbf{1}_\eta^T - \mathbf{C}_3 \otimes \mathbf{1}_\eta \right), \quad (30)$$

$$[\mathbf{K}]_{1:\eta,\eta+1:N} = -\frac{1}{2} \left([\mathbf{D}]_{1:\eta,\eta+1:N} + \mathbf{C}_1 \otimes \mathbf{1}_\eta \mathbf{1}_{N-\eta}^T - \mathbf{C}_2 \otimes \mathbf{1}_{N-\eta}^T - \mathbf{C}_4 \otimes \mathbf{1}_\eta \right), \quad (31)$$

where \otimes denotes the Kronecker product and

$$\mathbf{C}_1 = \frac{1}{\eta^2} \cdot \left[\mathbf{1}_\eta^T \cdot [\mathbf{D}]_{1:\eta,1:\eta} \cdot \mathbf{1}_\eta \right], \quad (32)$$

$$\mathbf{C}_2 = \frac{1}{\eta} \cdot \left[[\mathbf{D}]_{1:\eta,1:\eta} \cdot \mathbf{1}_\eta \right], \quad (33)$$

$$\mathbf{C}_3 = \frac{1}{\eta} \cdot \left[\mathbf{1}_\eta^T \cdot [\mathbf{D}]_{1:\eta,1:\eta} \right], \quad (34)$$

$$\mathbf{C}_4 = \frac{1}{\eta} \cdot \left[\mathbf{1}_\eta^T [\mathbf{D}]_{1:\eta,\eta+1:N} \right]. \quad (35)$$

Finally, invoke the relation (Dattorro, 2005, pp. 196)

$$\tilde{\mathbf{D}} = \left(\mathbf{1}_N \cdot \text{diag}(\tilde{\mathbf{K}})^T + \text{diag}(\tilde{\mathbf{K}}) \mathbf{1}_N^T - 2\tilde{\mathbf{K}} \right). \quad (36)$$

Equation 36 yields a complete set of distances associated with $\tilde{\mathbf{K}}$, such that any missing entries of $[\mathbf{D}]_{\eta+1:N,\eta+1:N}$ can be replaced by corresponding entries from $\tilde{\mathbf{D}}$.

At this point, let us emphasize that $[\mathbf{D}]_{1:\eta,1:\eta}$ contains the distances amongst anchors and consequently $[\mathbf{K}]_{1:\eta,1:\eta}$, \mathbf{C}_1 , \mathbf{C}_2 and \mathbf{C}_3 are all *constant*, such that $\tilde{\mathbf{K}}$ can be updated very efficiently.

Furthermore, the elements of $[\mathbf{D}]_{1:\eta,\eta+1:N}$ are the distances from anchors to targets, and therefore constitute the least (reasonable) amount of information required by tracking applications, such that this “completion” procedure can always² be applied.

In the extreme case of $[\mathbf{D}]_{\eta+1:N,\eta+1:N} = \mathbf{0}_{N-\eta}$ then to recover $[\mathbf{X}]_{\eta+1:N,1:\eta}$ only the eigendecomposition of $[\tilde{\mathbf{K}}]_{1:\eta,1:\eta}$ is required C. Fowlkes & Malik (2004).

Indeed, let $[\tilde{\mathbf{K}}]_{1:\eta,1:\eta} = \mathbf{Q} \cdot \mathbf{\Lambda} \cdot \mathbf{Q}^T$ as the eigendecomposition of $[\tilde{\mathbf{K}}]_{1:\eta,1:\eta}$. From equation 28 it follows that $[\mathbf{X}]_{1:\eta,1:\eta} = \mathbf{Q} \cdot \mathbf{\Lambda}^{\frac{1}{2}}$, and because $[\tilde{\mathbf{K}}]_{1:\eta,\eta+1:N} = [\mathbf{X}]_{1:\eta,1:\eta} \cdot [\mathbf{X}]_{\eta+1:N,1:\eta}^T$ then

$$[\mathbf{X}]_{\eta+1:N,:} = [\mathbf{X}]_{1:\eta,1:\eta}^T \cdot [\tilde{\mathbf{K}}]_{1:\eta,\eta+1:N} = \mathbf{Q} \cdot \mathbf{\Lambda}^{\frac{1}{2}} \cdot [\tilde{\mathbf{K}}]_{1:\eta,\eta+1:N}, \quad (37)$$

In conclusion, if an incomplete EDM is observed the aforementioned steps can be followed to complete \mathbf{D} before constructing the MDS kernel \mathbf{K}^* described in equations 27a.

² Even the case of sparse incomplete EDMs in which *none* of the rows of \mathbf{D} is complete could, in principle, also be dealt with by combining the Nystöm solution with standard completion algorithms applied to a restricted subset of η rows of \mathbf{D} Shang & Ruml (2004). This case, however, is of relatively little interest to tracking applications and outside the scope of the article.

5.2.3 SMACOF

The SMACOF technique is a well-known iterative algorithm that attempts to find the minimum of a non-convex function by tracking the global minima of the so-called majored convex functions $\mathcal{T}(\hat{\mathbf{X}}, \mathbf{Y})$ successively constructed from the original objective and basis on the previous solutions. In our context, the objective function to majorize is that one given in equation 7. Thus, we have

$$\mathcal{T}(\hat{\mathbf{X}}, \mathbf{Y}) = \sum w_{ij}^2 \cdot \bar{d}_{ij}^2 + \text{tr}\left(\hat{\mathbf{X}}^T \cdot \mathbf{H} \cdot \hat{\mathbf{X}}\right) - 2 \cdot \text{tr}\left(\hat{\mathbf{X}}^T \cdot \mathbf{A}(\mathbf{Y}) \cdot \mathbf{Y}\right), \quad (38)$$

where $\text{tr}(\cdot)$ denotes the trace, $\mathbf{Y} \in \mathbb{R}^{N \times \eta}$ is an auxiliary variable and the entries of \mathbf{H} and $\mathbf{A}(\mathbf{Y})$ are given by

$$h_{ij} = \begin{cases} \sum_{i=1}^N h_{ij}, & i = j, \\ i \neq j \\ -w_{ij}^2, & i \neq j, \end{cases} \quad (39a)$$

$$a_{ij} = \begin{cases} \sum_{i=1}^N a_{ij}, & i = j, \\ i \neq j \\ w_{ij}^2 \cdot \frac{\bar{d}_{ij}}{\|\mathbf{y}_i - \mathbf{y}_j\|_2}, & i \neq j, \end{cases} \quad (39b)$$

where $w_{ij} > 0$ if $e_{ij} \in E$ and $w_{ij} = 0$ otherwise.

At the ℓ -th iteration the global minimum $\hat{\mathbf{X}}^{(\ell)}$ of the majored function $\mathcal{T}(\hat{\mathbf{X}}, \mathbf{Y})$ with $\mathbf{Y} = \hat{\mathbf{X}}^{(\ell-1)}$, is computed via the Guttman transform,

$$\hat{\mathbf{X}}^{(\ell)} = \mathbf{H}^\dagger \cdot \mathbf{A}\left(\hat{\mathbf{X}}^{(\ell-1)}\right) \cdot \hat{\mathbf{X}}^{(n-1)}, \quad (40)$$

where \dagger denotes the pseudoinverse.

5.2.4 Linear global distance continuation

While the C-MDS and the Nyström approximation are algebraic approaches and SMACOF relies on the initialization point $\hat{\mathbf{X}}^{(0)}$, the algorithm proposed below, performs a low-complexity unconstrained global optimization. The approach is based on an iterative global smoothing technique, in which the global minimum is sought (with probability close to 1) after L number of iterations. The overall worse-case complexity of the method is equal to $L \times O$, where O is the worse-case complexity of the optimization technique used at the ℓ -th iteration. For convenience, we use a Quasi-Newton line search method whose complexity is $\mathcal{O}(N^2)$ Nocedal & Wright (2006). The proposed technique will be hereafter referred to as linear-global distance continuation (L-GDC) method, inspired by More & Wu (1997).

5.2.4.1 Fundamentals of the L-GDC

The objective of this subsection is to provide the fundamental Definitions, Theorems and Lemmas used in the L-GDC algorithm. Given the limited number of pages, we omit all proofs which can be found in Destino & Abreu (2010); More & Wu (1997).

Definition 1 (Gaussian kernel) Let $g(u, \lambda)$ be the Gaussian kernel defined as

$$g(u, \lambda) \triangleq e^{-u^2/\lambda^2}. \quad (41)$$

Definition 2 (Gaussian transform) Let $\langle s \rangle_\lambda(\mathbf{x})$ denote the Gaussian transform (smoothed function) of a function $s(\mathbf{x})$, and given by

$$\langle s \rangle_\lambda(\mathbf{x}) \triangleq \frac{1}{\pi^{n/2} \lambda^n} \int_{\mathbb{R}^n} s(\mathbf{u}) e^{-\frac{\|\mathbf{x}-\mathbf{u}\|_F^2}{\lambda^2}} d\mathbf{u}, \quad (42)$$

where $\mathbf{u}, \mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^+$ is a parameter that controls the degree of smoothing ($\lambda \gg 0$ strong smoothing).

Theorem 1 (Continuation method) Let $\{\lambda^{(\ell)}\}$ with $\{1 \leq \ell \leq L\}$ be any sequence of λ 's converging to zero, i.e. $\lambda^{(L)} = 0$. If $\mathbf{x}^{(\ell)}$ is a global minimizer of $\langle s \rangle_{\lambda^{(\ell)}}(\mathbf{x})$ and $\{\mathbf{x}^{(\ell)}\}$ converges to \mathbf{x}^* , then \mathbf{x}^* is a global minimizer of $\langle s \rangle_{\lambda^{(0)}}(\mathbf{x})$.

Theorem 2 (WLS-ML Smoothed Function) Let $s(\hat{\mathbf{X}})$ equal to the objective function in equation 7, and for simplicity consider $\eta = 2$. Then, the smoothed function $\langle s \rangle_\lambda(\hat{\mathbf{X}})$ is given by

$$\begin{aligned} \langle s \rangle_\lambda(\hat{\mathbf{X}}) &= \frac{1}{\pi} \int_{\mathbb{R}^\eta} \sum_{ij} w_{ij}^2 (\tilde{d}_{ij} - \hat{d}_{ij,u})^2 \exp(\|\mathbf{u}\|_F^2) d\mathbf{u}, \\ &= \sum_{ij} w_{ij}^2 \left(\lambda^2 + \tilde{d}_{ij}^2 + \hat{d}_{ij}^2 - 2\lambda \tilde{d}_{ij} \Gamma\left(\frac{3}{2}\right) {}_1F_1\left(\frac{3}{2}; 1; \frac{\hat{d}_{ij}^2}{\lambda^2}\right) \exp\left(\frac{-\hat{d}_{ij}^2}{\lambda^2}\right) \right), \end{aligned} \quad (43)$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{N \times \eta}$ is a matrix whose i -th row-vector is $\hat{\mathbf{x}}_i$, $\hat{d}_{ij,u} \triangleq \|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j + \lambda \mathbf{u}\|_F$ with $\mathbf{u} \in \mathbb{R}^\eta$, $\Gamma(a)$ is the gamma function and ${}_1F_1(a; b; c)$ is the confluent hypergeometric function Abramowitz & Stegun (1965).

Theorem 3 (Convexity condition) Let $s(\hat{\mathbf{X}})$ equal to the objective function in equation 7 and $\langle s \rangle_\lambda$ given by 43, then $\langle s \rangle_\lambda$ is convex if

$$\lambda^* \geq \frac{\sqrt{\pi} \max_{ij} \tilde{d}_{ij}}{2}. \quad (44)$$

Lemma 1 (Minimal $\{\lambda^{(\ell)}\}$ for source localization)

Let $N_T = 1$ and let $\hat{\mathbf{x}}$ denote the target location estimate. Consider an ordered set of ranging measurement $\{\tilde{d}_\ell\}$, such that $\tilde{d}_1 \geq \tilde{d}_2 \geq \dots \tilde{d}_L$, where $L = N_A$. Then the minimal set $\{\lambda^{(\ell)}\}$ is given by

$$\lambda^{(\ell)} = \frac{\sqrt{\pi} \tilde{d}_\ell}{2}, k = 1 \dots L. \quad (45)$$

5.2.4.2 Implementation of the L-GDC

Invoking Theorems 1,2 and 3 the L-GDC algorithm is given by

$$\mathbf{x}^{(\ell)} = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \langle s \rangle_{\lambda^{(\ell)}}(\mathbf{x}), 1 \leq \ell \leq L, \quad (46)$$

with

$$\lambda_0 = \frac{\sqrt{\pi} \max_{ij} \bar{d}_{ij}}{2}. \quad (47)$$

Notice that the ML-transformed objective function involves the hypergeometric function then the numeric evaluation of equation 43 requires care, especially when λ is very small. Numerically stable computations can be achieved using the equivalences Abramowitz & Stegun (1965)

$${}_1F_1\left(\frac{3}{2}; 1; z\right) = 1 + \sum_{m=1}^{+\infty} \left(z^m \cdot \prod_{k=1}^m \frac{(1/2+k)}{k^2} \right), z < 10. \quad (48)$$

$${}_1F_1\left(\frac{3}{2}; 1; z\right) \approx \frac{z^{-3/2}}{\Gamma(-\frac{1}{2})} \left(\sum_{m=0}^{M-1} \frac{(-z)^{-m} m^{-1}}{m!} \prod_{k=0}^{m-1} \left(\frac{3}{2} + k \right)^2 \right) + \frac{e^z z^{1/2}}{\Gamma(\frac{3}{2})} \left(\sum_{p=0}^{P-1} \frac{z^{-p} p^{-1}}{p!} \prod_{k=0}^{p-1} \left(k - \frac{1}{2} \right)^2 \right), z \geq 10, \quad (49)$$

where $z \triangleq \frac{\hat{d}^2}{\lambda^2}$ and M and P are sufficiently large numbers to ensure an accurate approximation (typically $(M, P) \geq 5$).

In order to use a Newton's based optimization method, gradient and Hessian of $\langle s \rangle_{\lambda^{(t)}}(\mathbf{x})$ are required. The gradient is given by

$$\nabla_{\hat{\mathbf{x}}} \langle s \rangle_{\lambda}(\hat{\mathbf{X}}) = \sum_{ij} w_{ij}^2 s'_{ij}(\hat{d}_{ij}; \lambda) \nabla_{\hat{\mathbf{x}}}(\hat{d}_{ij}), \quad (50)$$

where the i -th and the j -th $1 \times \eta$ blocks of $\nabla_{\hat{\mathbf{x}}}(\hat{d}_{ij})$ are

$$\left[\nabla_{\hat{\mathbf{x}}}(\hat{d}_{ij}) \right]_i = \frac{\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_i}{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_2}, \quad (51)$$

$$\left[\nabla_{\hat{\mathbf{x}}}(\hat{d}_{ij}) \right]_j = -\frac{\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_i}{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_2}, \quad (52)$$

and the function $s'_{ij}(\hat{d}_{ij}; \lambda)$ is the first derivative of

$$s_{ij}(\hat{d}_{ij}; \lambda) \triangleq \bar{d}_{ij}^2 + \hat{d}_{ij}^2 + \lambda^2 - \lambda \hat{d}_{ij} \sqrt{\pi} {}_1F_1\left(\frac{3}{2}, 1, \frac{\hat{d}_{ij}^2}{\lambda^2}\right) \exp\left(\frac{-\hat{d}_{ij}^2}{\lambda^2}\right), \quad (53)$$

and is equal to

$$s'_{ij}(\hat{d}_{ij}; \lambda) = 2\hat{d}_{ij} + \frac{\hat{d}_{ij} \sqrt{\pi} \bar{d}_{ij}}{\lambda} S_1(\hat{d}_{ij}; \lambda), \quad (54)$$

where

$$S_1(\hat{d}_{ij}; \lambda) \triangleq \exp\left(\frac{-\hat{d}_{ij}^2}{\lambda^2}\right) \left(2 {}_1F_1\left(\frac{3}{2}; 1; \frac{\hat{d}_{ij}^2}{\lambda^2}\right) - 3 {}_1F_1\left(\frac{5}{2}; 2; \frac{\hat{d}_{ij}^2}{\lambda^2}\right) \right). \quad (55)$$

The Hessian matrix of $\langle s \rangle_\lambda(\hat{\mathbf{X}})$, denoted by $\nabla_{\hat{\mathbf{X}}}^2 \langle s \rangle_\lambda(\hat{\mathbf{X}})$, is computed as

$$\nabla_{\hat{\mathbf{X}}}^2 \langle s \rangle_\lambda(\hat{\mathbf{X}}) = \sum_{ij} w_{ij}^2 \left(s_{ij}''(\hat{d}_{ij}; \lambda) \nabla_{\hat{\mathbf{X}}}^T(\hat{d}_{ij}) \nabla_{\hat{\mathbf{X}}}(\hat{d}_{ij}) + s'_{ij} \nabla_{\hat{\mathbf{X}}}^2(\hat{d}_{ij}) \right), \quad (56)$$

where $\nabla_{\hat{\mathbf{X}}}^2(\hat{d}_{ij}) \in \mathbb{R}^{N\eta \times N\eta}$ is given by a symmetric block-matrix where the ii -th and ij -th blocks are

$$\left[\nabla_{\hat{\mathbf{X}}}^2(\hat{d}_{ij}) \right]_{ii} = \frac{1}{\hat{d}_{ij}} \left(\mathbf{I} - \left[\nabla_{\hat{\mathbf{X}}}(\hat{d}_{ij}) \right]_i^T \left[\nabla_{\hat{\mathbf{X}}}^2(\hat{d}_{ij}) \right]_i \right), \quad (57)$$

$$\left[\nabla_{\hat{\mathbf{X}}}^2(\hat{d}_{ij}) \right]_{ij} = - \left[\nabla_{\hat{\mathbf{X}}}(\hat{d}_{ij}) \right]_{ii}, \quad (58)$$

and the second derivative of $s_{ij}(\hat{d}_{ij}; \lambda)$, denoted by $s_{ij}''(\hat{d}_{ij}; \lambda)$, is

$$s_{ij}''(\hat{d}_{ij}; \lambda) = 2 + \frac{\sqrt{\pi} \bar{d}_{ij}}{\lambda} S_1(\hat{d}_{ij}; \lambda) + \frac{\sqrt{\pi} \bar{d}_{ij} \hat{d}_{ij}}{\lambda^3} (S_2(\hat{d}_{ij}; \lambda) - S_1(\hat{d}_{ij}; \lambda)), \quad (59)$$

with

$$S_2(\hat{d}_i; \lambda) \triangleq e^{-\frac{\hat{d}_i^2}{\lambda^2}} \left(3 {}_1F_1 \left(\frac{5}{2}; 2; \frac{\hat{d}_i^2}{\lambda^2} \right) - \frac{15}{4} {}_1F_1 \left(\frac{7}{2}; 3; \frac{\hat{d}_i^2}{\lambda^2} \right) \right). \quad (60)$$

In what follows, we provide an example of source localization problem in $\eta = 1$ dimension. Let $N_A = 2$ and $N_T = 1$. The anchors' and target coordinates are $\mathbf{a}_1 = 0.2$, $\mathbf{a}_2 = 0.5$ and $\mathbf{x} = 0.5$, respectively. We assume no noise, thus $\tilde{d}_i = d_i$. Invoking Theorem 3 we compute $\lambda^{(0)} = 0.3988$ such that $\langle s \rangle_\lambda(\hat{x})$ is convex. Next, we apply the L-GDC technique summarized in equation 46 where, invoking Theorem 2, we choose the set of λ 's such that $\lambda^{(0)} = 0.3988$, $\lambda^{(L)} = 0$ and $\lambda^{(\ell)} = \lambda^{(\ell-1)} - 0.05$.

The light-gray lines shown in figure 8 indicate the smoothed function computed with the aforementioned set. At each iteration the level of smoothing is decreased and $\langle s \rangle_\lambda(\hat{x})$ approaches more and more $s(\hat{x})$.

The bold lines correspond, instead, to the smoothed function computed for the set of λ 's $\{0.3988, 0.3420, 0.1706, 0\}$ using the Lemma 1. In this case, it is shown that $\langle s \rangle_\lambda(\hat{x})$ is recomputed only when a new concave region appears, thus we drastically reduce the computational efficiency of the L-GDC method while preserving optimal performance.

6. Simulation results

In this section, the performance of the non-parametric WLS-ML LT approach considered in this chapter will be evaluated using different optimization algorithms and adopting different weighing strategies. We will use the root-mean-square-error (RMSE) to measure the accuracy of the estimated positions $\hat{\mathbf{X}}$

$$\text{RMSE} \triangleq \sqrt{\frac{1}{RP} \sum_{p=1}^P \sum_{r=1}^R \|\hat{\mathbf{X}}_{rp} - \mathbf{X}\|_2^2}, \quad (61)$$

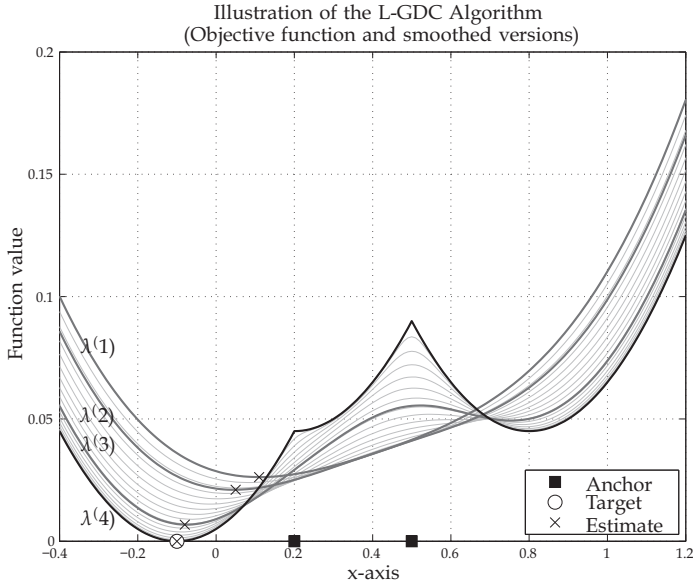


Fig. 8. Illustration of the L-GDC method and the smoothing process. Light-gray lines indicate smoothed versions of the objective functions obtained with a linear decreasing sequence of λ 's. Bold lines indicate the smoothed objective with the optimized λ selection criteria.

where R and P are respectively, the number of realizations and networks considered. For the purpose of comparison, we also benchmark the results to the Cramér-Rao lower bound (CRLB) derived in Jourdan et al. (2006) Patwari et al. (2003) and given by

$$\text{CRLB} \triangleq \text{tr}(\mathbf{F}^\dagger), \tag{62}$$

where \mathbf{F} is the Fisher information matrix, that for $\eta = 2$ is equal to

$$\mathbf{F} \triangleq \begin{bmatrix} \mathbf{F}_{xx} & \mathbf{F}_{xy} \\ \mathbf{F}_{xy}^T & \mathbf{F}_{yy} \end{bmatrix}, \tag{63}$$

where

$$[\mathbf{F}_{xx}]_{jl} = \begin{cases} \sum_{e_j \in E} \frac{K_{il}}{\sigma_{il}^2} \frac{(x_l - x_i)^2}{d_{il}^2}, & j = l \\ -\frac{K_{il}}{\sigma_{il}^2} \frac{(x_l - x_j)^2}{d_{jl}^2}, & j \neq l \text{ and } e_j \in E \end{cases} \tag{64}$$

$$[\mathbf{F}_{yy}]_{jl} = \begin{cases} \sum_{e_j \in E} \frac{K_{il}}{\sigma_{il}^2} \frac{(y_l - y_i)^2}{d_{il}^2}, & j = l \\ -\frac{K_{il}}{\sigma_{il}^2} \frac{(y_l - y_j)^2}{d_{jl}^2}, & j \neq l \text{ and } e_j \in E \end{cases} \tag{65}$$

$$\left[\mathbf{F}_{xy} \right]_{jl} = \begin{cases} \sum_{e_j \in E} \frac{K_{il}}{\sigma_{il}^2} \frac{(x_l - x_i)(y_l - y_i)}{d_{il}^2}, & j = l \\ -\frac{K_{il}}{\sigma_{il}^2} \frac{(x_l - x_j)(y_l - y_j)}{d_{jl}^2}, & j \neq l \text{ and } e_{jl} \in E \end{cases} \tag{66}$$

where e_j indicates the set of links connected to the j -th node.

The first case-of-study is a network with $N_A = 4$ anchors and one target deployed in a square area of size $[-10,10] \times [-10,10]$. The target location is generated as a random variable with uniform distribution within the size of the square while anchors, are located at the locations $\mathbf{x}_1 = [-10,-10]$, $\mathbf{x}_2 = [10,-10]$, $\mathbf{x}_3 = [10,10]$ and $\mathbf{x}_4 = [-10,10]$. We assume that all nodes are connected and the distance of each link is measured K_{ij} times, with $K_{ij} \in [2,7]$. We use the ranging model given in equation 2 to generate distance measurements, and we consider $\sigma_{ij} \in (1e-4, \sigma_{\max})$.

In figure 9, we show the RMSE obtained with different localization algorithms and unitary weight (unweighted strategy). In this particular study, all algorithms have very similar performance, and the reason is due to the convexity property of the WLS-ML objective function. Indeed, if the target is inside the convex-hull formed by the anchors and the noise is not sufficiently large, then the objective function is typically convex. However, all algorithms do not attain the CRLB because, under the assumption that σ_{ij} 's are all different, the unitary weight is not optimal.

In figure 10 we show the RMSE obtained with the L-GDC algorithm using different weighing strategy, namely, the optimal, the unweighted, the exponential and the dispersion weighing strategy given in equations12, 14, 17, and20, respectively. The results show that the L-GDC algorithm using w_{ij}^* is able to achieve the CRLB, whereas the others stay above.

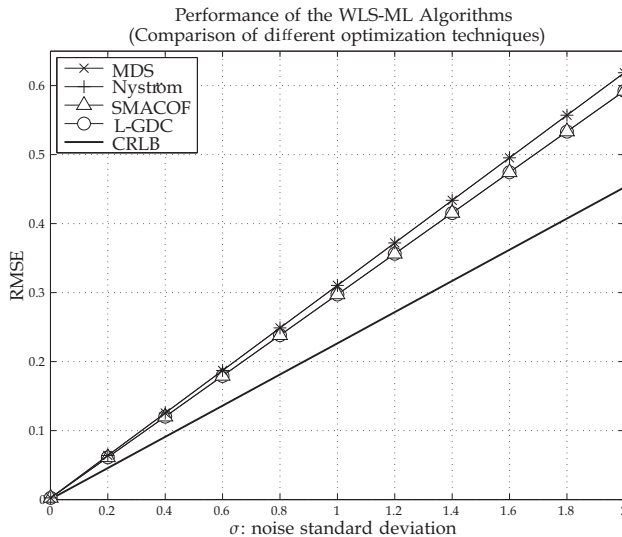


Fig. 9. Comparison of different optimization techniques and using binary weight (unweighted strategy) for a localization problem with $N_A = 4$, $N_T = 1$, $K_{\min} = 2$, $K_{\max} = 7$, $\sigma_{\max} = 1$ and $\sigma_{\min} = 1e-4$.

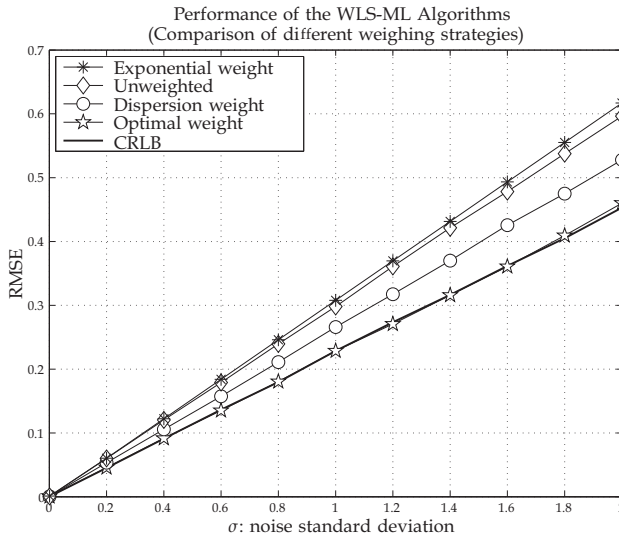


Fig. 10. Comparison of different weighing strategies and using L-GDC optimization method for a localization problem with $N_A = 4$, $N_T = 1$, $K_{\min} = 2$, $K_{\max} = 7$, $\sigma_{\max} = 1$ and $\sigma_{\min} = 1e-4$.

However, to use the optimal weighing strategy we assumed that σ_{ij} 's are known a priori. Therefore, if we reconsider the LT problem under the assumption that the noise statistics are unknown, then the proposed dispersion weight provides the best performance. Indeed, using w_{ij}^l we are able to rip $\approx 50\%$ of gain from the unweighted and exponential strategies towards the CRLB.

In the second case-of-study, we consider instead a network with $N_A = 4$ anchors and $N_T = 10$ targets. As before, anchors are located at the corners of a square area while targets are randomly distributed. For this type of simulations, we evaluate the performance of the WLSML algorithms as functions of the *meshness ratio* defined as

$$m \triangleq \frac{(|E| - N + 1)}{(|E_F| - N + 1)}, \quad (67)$$

where E_F indicates the set of links of the fully connected network and $|\cdot|$ indicates the cardinal number of a set Adams & Franzosa (2008)Destino & De Abreu (2009).

This metric is commonly used in algebraic topology and Graph theory to capture, in one number, information on the planarity of a Graph. For example, under the constraint of a connected network, $m = 0$ results from $|E| = N - 1$, which implies that the network is reduced to a tree. In contrast, $m = 1$ results from $|E| = |E_F|$, which implies that the network is not planar, except for the trivial cases of $N \leq 4$. More importantly, the meshness ratio is an indicator of the connectivity of the network, in a way that is more relevant to its localizability than the simpler connectivity ratio $|E| / |E_F|$.

In figures 11 and 12, the results confirm that the L-GDC is the best optimization technique and, the dispersion weight is the best performing weighing strategy. Similarly to the first case-of-study, also in this case the WLS-ML method based on L-GDC and using the dispersion weights rips about 50% of the error from the alternatives towards the CRLB. Furthermore, from the results shown in figure 11, the L-GDC algorithm is the only one to

maintain an almost constant gap from the CRLB within the entire range of meshness ratio. This let us infer that the L-GDC algorithm finds the global optimum of the WLS-ML function with high probability, while SMACOF of the algebraic methods find sub-optimal solutions.

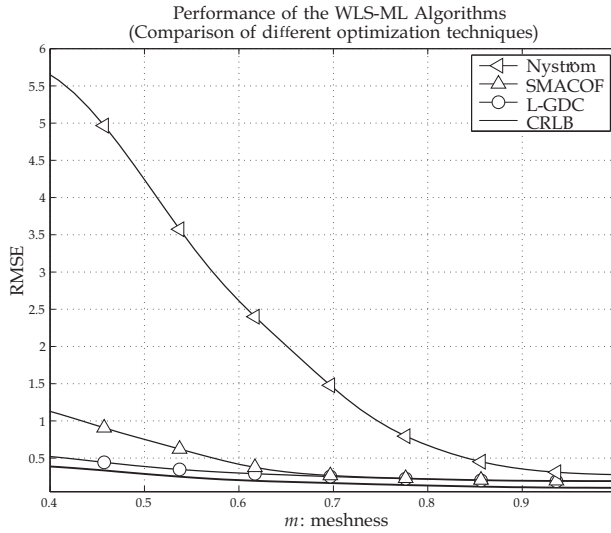


Fig. 11. Comparison of different optimization techniques and using binary weight (unweighted strategy) for a localization problem with $N_A = 4$, $N_T = 10$, $K_{min} = 2$, $K_{max} = 7$, $\sigma_{max} = 1$ and $\sigma_{min} = 1e-4$.

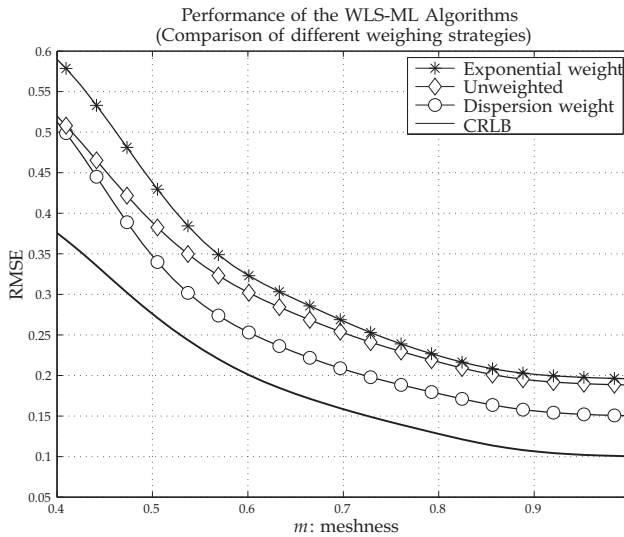


Fig. 12. Comparison of different weighing strategies and using L-GDC optimization method for a localization problem with $N_A = 4$, $N_T = 10$, $K_{min} = 2$, $K_{max} = 7$, $\sigma_{max} = 1$ and $\sigma_{min} = 1e-4$.

The third and final case-of-study, is the tracking scenario. The network consists of 4 anchor nodes placed at the corner of a square in a $\eta = 2$ dimensional space with 1 targets that moves following an autoregressive model of order 1 within space defined by the anchors. It is assumed full anchor-to-anchor and anchor-to-target connectivity and measurements are perturbed by zero-mean Gaussian noise.

We use the L-GDC optimization method to perform successive re-localization of the target and we employ different weighing strategies. The result shown in figure 14 illustrates the performance of the WLS-ML algorithm as a function of σ considering a velocity $v = 1$.

Since the tracking is treated as a mere re-localization, the dynamics only affect the output of the filter block and it is seen from the localization algorithm as an additive noise.

For this reason, the trend of the RMSE is similar to that one obtained in a static scenario. From figure 14 the impact of the velocity on the performance of the WLS-ML algorithm with wavelet-based filter is revealed more clearly. The effect of velocity, indeed, is yet similar to a gaussian noise.

Finally, from both results we observe that the dispersion weight is the best weighing strategy.

7. Conclusions and future work

In this chapter we considered the LT problem in mesh network topologies under LOS conditions. After a general description of the system we focused on a wavelet based filter to smooth the observations and a centralized optimization technique to solve the WLS-ML localization problem. The proposed algorithm was compared with state-of-the-art solutions and it was shown that by combining the wavelet-based filter together with the dispersion weighing strategy and the L-GDC algorithm it is possible to get close to the CRLB.

The work described in this chapter did not address the problem of NLOS channel conditions which needs to be taken into consideration in most of the real life applications. To cope with the biases introduced by NLOS condition two main strategies can be distinguished. In the

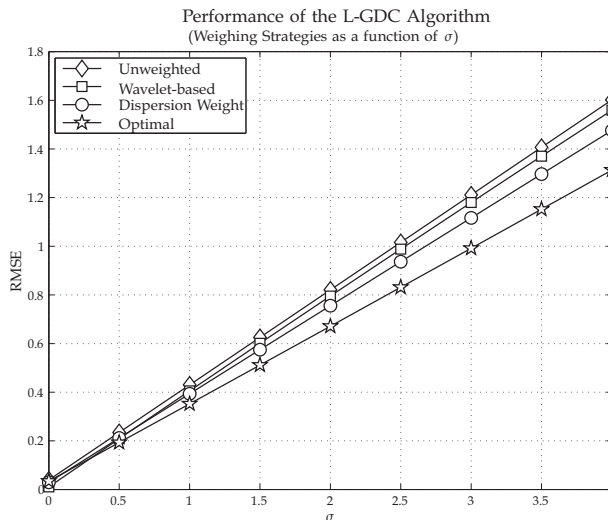


Fig. 13. Performance for the L-GDC algorithm for the different weighing strategies. Scenario measurements at the 4 anchor nodes subject to normal noise process with standard deviation between 0 and σ .

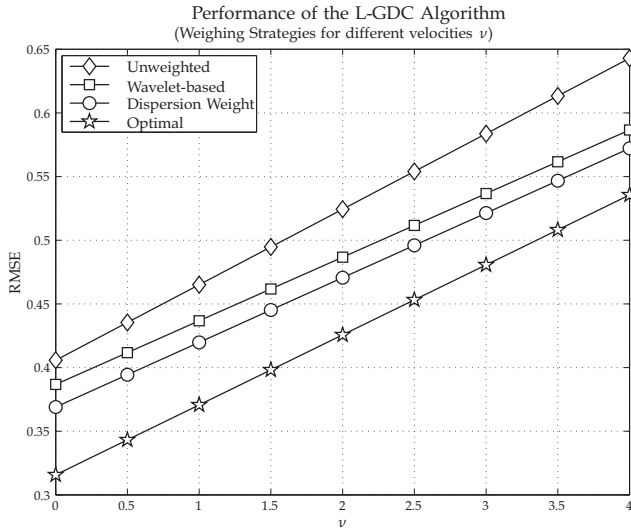


Fig. 14. Performance for the L-GDC algorithm for the different weighing strategies. Scenario measurements at the 4 anchor nodes subject to normal noise process with $\sigma = 2$ and variable target dynamic ν .

first one the biases are treated as additional variables and are directly estimated by the LT algorithm while the second approach aims at discarding the bias introduced by the NLOS condition by applying channel identification and bias compensation algorithms before the LT engine. Concluding, a new method recently proposed by the authors to overcome the NLOS effects is based on an accurate contraction of all the measured distances which has been shown to positively affect the convexity of the objective function and consequently the final location estimates.

8. References

- Abramowitz, M. & Stegun, I. A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10 edn, Dover Publications.
- Adams, C. & Franzosa, R. (2008). *Introduction to Topology Pure and Applied*, Pearson Prentice Hall.
- Alfakih, A. Y., Wolkowicz, H. & Khandani, A. (1999). Solving euclidean distance matrix completion problems via semidefinite programming, *Journ. on Comp. Opt. and App.* 12(1): 13 – 30.
- Beck, A., Stoica, P. & Li, J. (2008). Exact and approximate solutions for source localization problems, *IEEE Trans. Signal Processing* 56(5): 1770–1778.
- Biswas, P., Liang, T.-C., Toh, K.-C. & Wang, T.-C. (2006). Semidefinite programming based algorithms for sensor network localization with noisy distance measurements, *ACM Trans. on Sensor Netw. (TOSN)* 2(2): 188–220.
- Biswas, P., Liang, T.-C., Toh, K.-C., Wang, T.-C. & Ye, Y. (2006). Semidefinite programming approaches for sensor network localization with noisy distance measurements, *IEEE Trans. Autom. Sci. Eng.* 3: 360–371.
- Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press.
- C. Fowlkes, S. Belongie, F. C. & Malik, J. (2004). Spectral grouping using the Nyström method, *IEEE Trans. Pattern Anal. Machine Intell.* 26(2).

- Cheung, K., So, H., Ma, W.-K. & Chan, Y. (2004). Least squares algorithms for time-of-arrival-based mobile location, *IEEE Trans. on Signal Processing* 52(4): 1121-1130.
- Costa, J. A., Patwari, N. & III, A. O. H. (2006). Distributed multidimensional scaling with adaptive weighting for node localization in sensor networks, *ACM J. on Sensor Netw.* 2(1): 39-64.
- Cox, T. F. & Cox, M. A. A. (2000). *Multidimensional Scaling*, 2 edn, Chapman & Hall/CRC.
- Dattorro, J. (2005). *Convex Optimization and Euclidean Distance Geometry*, Meboo Publishing.
- Destino, G. & Abreu, G. (2009). Solving the source localization problem via global distance continuation, *Proc. IEEE International Conference on Communications*. IEEE Asilomar Conference on Signals, Systems, and Computers.
- Destino, G. & Abreu, G. (2010). On the maximum likelihood formulation of the network localization problem, (to submit).
- Destino, G. & De Abreu, G. T. F. (2009). Weighing strategy for network localization under scarce ranging information, *Trans. Wireless. Comm.* 8(7): 3668-3678.
- Gibbons, J. (1992). *Nonparametric Statistical Inference*, Marcel Dekker.
- Guvenc, I., Gezici, S., Watanabe, F. & Inamura, H. (2008). Enhancements to linear least squares localization through reference selection and ML estimation, *Proc. IEEE Wireless Comm. and Netw. Conf. (WCNC)*, pp. 284-289.
- Joon-Yong, L. & Scholtz, R. (2002). Ranging in a dense multipath environment using an UWB radio link., *IEEE J. Sel. Areas Commun.* 20: 1667-1683.
- Jourdan, D., Dardari, D. & Win, M. (2006). Position error bound for UWB localization in dense cluttered environments, *Proc. IEEE International Conference on Communications*, Vol. 8, pp. 3705-3710.
- Li, X. & Pahlavan, K. (2004). Super-resolution toa estimation with diversity for indoor geolocation, *IEEE Trans. Wireless Commun.* 3(1): 224-234.
- Macagnano, D. & de Abreu, G. T. F. (2008). Tracking multiple dynamic targets in LOS-NLOS condition with multidimensional scaling, *IEEE 5th Workshop on Positioning, Navigation and Communication*.
- Mao, G., Fidan, B. & Anderson, B. D. O. (2007). Wireless sensor network localization techniques, *Computer Networks: The Intern. J. of Comp. and Telecomm. Networking* 51(10): 2529-2553.
- More, J. & Wu, Z. (1997). Global continuation for distance geometry problems, *SIAM J. Optim.* 7: 814-836.
- Nocedal, J. & Wright, S. (2006). *Numerical Optimization*, Springer.
- Ouyang, R., Wong, A.-S. & Chin-Tau, L. (2010). Received signal strength-based wireless localization via semidefinite programming: Noncooperative and cooperative schemes, *IEEE Transactions on Vehicular Technology* 59(3): 1307 -1318.
- Patwari, N., Dea, R. J. O. & Wang, Y. (2003). Relative location estimation in wireless sensor networks, *IEEE Trans. Signal Processing* 51(8): 2137-2148.
- Shang, Y. & Ruml, W. (2004). Improved MDS-based localization, *Proc. 23-rd Ann. Joint Conf. of the IEEE Comp. and Comm. Societies (INFOCOM'04)*, Vol. 4, Hong-Kong, China, pp. 2640 - 2651.
- S.Mallat (1998). *A Wavelet Tour of Signal Processing*, second edn, Academic Press.
- S.Mallat & S.Zhong (1992). Characterization of signals from multiscale edges, *IEEE Trans. Pattern Anal. Machine Intell.* 14(7): 710-732.
- Williams, C. & M.Seeger (2000). Using the Nyström method to speed up kernel machines, *Annual Advances in Neural Information Processing Systems* 13 pp. 682-688.

Usage of Mesh Networking in a Continuous-Global Positioning System Array for Tectonic Monitoring

Hoang-Ha Tran and Kai-Juan Wong
*Nanyang Technological University
Singapore*

1. Introduction

In recent years, tectonic plate movements have caused huge natural disasters, such as the Great Sumatra-Andaman earthquake and the resulting Asian tsunami, which led to significant loss of human lives and properties (Ammon et al., 2005; Lay et al., 2005). Scientific evidences proved it was the beginning of a new earthquake supper-cycle in this active area (Sieh et al., 2008). In order for scientists to further study such disasters and provide early warning of imminent seismic events, many continuous-Global Positioning System (cGPS) arrays were developed and deployed to monitor the active tectonic plates around the world such as “SuGAR” along the Sumatran fault, “GEONET” covering all Japan islands, and “SCIGN” covering most of southern California. Each of these cGPS arrays contains tens to hundreds of GPS stations. Using precise GPS receivers, antennas and scientific-grade GPS processing software, measurements from each GPS station are able to provide location information with sub-millimeter accuracy. These location data produced by the GPS stations, which are located in the vicinity of active tectonic plates, provided accurate measurements of tectonic movements during the short period of a co-seismic event as well as for the long period observation of post-seismic displacement.

The GPS applications in earthquake studies (Segall & Davis, 1997) include monitoring of co-seismic deformation, post seismic and inter-seismic processes. Post seismic (except aftershocks) and inter-seismic deformations are much smaller than co-seismic events, where there is little or no supporting information from seismic measurements. In this instance, GPS can be used to detect the long time inter-seismic strain accumulation which leads to indentify the location of future earthquake (Konca et al., 2008).

In cGPS arrays utilizing satellite communications such as the Sumatran cGPS Array (SuGAR), each GPS station in the cGPS array will periodically measure the tectonic and/or meteorological data which will be stored locally. A collection of these observed GPS data will then be sent to a data server through a dedicated satellite link from each station either in real-time or at update intervals ranging from hours to months. At the server, the collected data from the GPS stations will be processed by using closely correlated data from each station to reduce errors in the location measurements. Since the amount of data transmitted from each station could be relatively large, the communication bandwidth and the number of uplinks are the most important factors in terms of operational expenditure. Each satellite

link requires costly subscription and data transmission across these links are usually charged based on the connection time or the amount of data transmitted/received. Therefore, in order to reduce the operational cost of a cGPS array, it is paramount that the number of satellite links as well as the data sent on these links be kept to a minimum. The rest of this chapter is organized as follows. Commonly used data formats for GPS processing is introduced in section 2. Introduction of cGPS arrays including SuGAR are presented in section 3. Proposed modifications of SuGAR network and parallel GPS processing which make use of mesh network are evaluated in section 4. Lastly, the chapter will end with a brief conclusion.

2. Common data formats used for cGPS systems

Scientific-grade GPS receivers store their measured signals in binary format that prolong logging time of those devices. Some of the most commonly used property binary formats for GPS receivers are R00/T00/T01/T02 and B-file/E-file used by Trimble and Ashtech receivers respectively. Another widely adopted binary format proposed by UNAVCO is the "BINary EXchange" (BINEX) format, which is used for research purposes. It has been designed to encapsulate most of the information currently acceptable for GPS data. Binary files were converted to text file for easy handling and processing. For GPS data storage and transmission, the most generally used GPS exchange data type is the RINEX format (Gurtner & Mader, 1990). It contains processed data collected by the GPS stations. This format defined four file types for observation data, navigation message, meteorological message and GLONASS navigation message. As correlation exists between the consecutive GPS measurement data, CRINEX (Hatanaka, 1996), a compressed RINEX format, proposed based on the idea that observation information between each measurement was related and changed at a small pace. The use of CRINEX reduces the storage space and transmission bandwidth requirements as only the difference between the current observation data and the first occurrence of it is stored.

3. Sumatran cGPS array - introduction and configuration

Many cGPS arrays were deployed to monitor some of the active tectonic plates around the world. Each of these cGPS arrays contains tens to hundreds of GPS stations, spanning from hundreds to thousands kilometers and varying methods are used for monitoring and harvesting the data from those stations. In this section, some of those arrays are described.

The GPS Observation Network system (GEONET) (Yamagiwa et al., 2006) is one of the most dense cGPS network comprising of over 1200 GPS stations nationwide. It was used to support real-time crustal deformation monitoring and location-based services. GEONET provides real-time 1Hz data through a dedicated IP-VPN (Internet Protocol Virtual Private Network).

The Southern California Integrated GPS Network (SCIGN) (Hudnut et al., 2001) contain more than 250 stations covering most of southern California which provide near real-time GPS data. SCIGN is used for fault interaction and post-seismic deformation in the eastern California shear zone.

The New Zealand GeoNet (Patterson et al., 2007) is a nation-wide network of broadband and strong ground motion seismometers complimented by regional short period seismometers and cGPS stations, volcano-chemical analyzers and remote monitoring

capabilities. It comprises of more than 150 cGPS stations across New Zealand. All seismic and GPS data are transmitted continuously to two data centers using radio, land-based or VSAT systems employing Internet Protocol data transfer techniques.

The Sumatran continuous-Global Positioning System Array (SuGAR) is located along Sumatra, Indonesia. As at the end of 2009, it consists of 32 operational GPS stations spanning 1400 km from north to south of Sumatra (Fig. 1). Stations are located either in remote islands or in rural areas near the tectonic place boundary which is one of the most active plates in the world. Due to the lack of local data communication network infrastructure, satellite telemetry is the only means of communicating with the GPS stations. All of the stations are equipped with a scientific-grade GPS receiver, a GPS antenna, a satellite modem, solar panels and batteries.

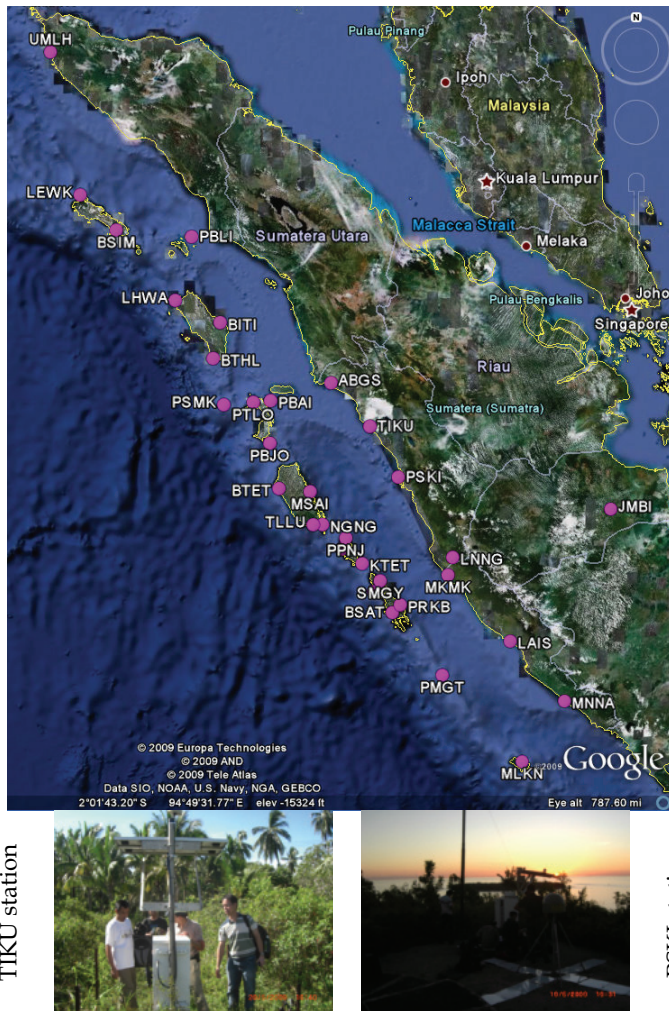


Fig. 1. Geographical distribution of the SuGAR stations

4. Utilisation of mesh networking

Mesh networking is proposed in this chapter to reduce the number of satellite links and bandwidth requirement for transmission of GPS data. To analyze the optimization achieved by the use of mesh networking on the SuGAR network, evaluation was performed using the archived SuGAR observation data from the last two months (61 days) of 2007. Only 24 stations were taken into account in this case study, as only 24 GPS stations were able to provide the complete GPS dataset for this entire period. This experiment data set can be accessed from the SOPAC website (<http://sopac.ucsd.edu/>).

Several assumptions were made for the evaluations presented in this study as follows:

- All GPS stations have enough energy to deal with the overheads cause by the additional communication equipments and data computation required. This assumption can be satisfied by adding more batteries and solar panels to the existing nodes.
- To simplify the analysis, the terrain information between the GPS stations was not taken into consideration in this analysis. In practice, construction of tall antenna towers as well as the use of multi-hop relays/repeaters can be used to overcome obstructions if required.
- The transmission overheads for the long range radios, such as packet formatting and control protocols, were not included in the evaluation as they will not have an impact on the analysis presented in this study.

The two main performance attributes of interest in this study are the reduction of the number of satellite links as well as the total amount of data transmitted via these links.

4.1 Removal of co-related data and reduction of uplink requirements

Mesh networking and clustering can be used to reduce the number of satellite links required for data telemetry between the GPS stations and the remote server. Wireless mesh networks can be established using long-range radios such as those developed by companies like FreeWave or Intuicom. These radios provide a point-to-point line-of-sight (LoS) wireless communication link with a maximum range of more than 96 kilometres (60 miles) and a maximum over-the-air throughput of 154 Kbps. For communication links over a longer distance, multi-hop communications can be utilized by deploying relay stations. The use of relay stations may also overcome LoS obstructions between GPS stations as well as provide for extended mesh networking capabilities such as redundancy. Depending on the cost, geographical, power or latency considerations, the number of hops and the radio range supported may be limited. In this case, clusters of GPS stations will be formed and a cluster-head would be selected for each cluster. Each cluster-head will have satellite communication capabilities and will be responsible for collecting all the observation data from the GPS stations within the cluster and transmitting them to the remote centralized data server. This greatly reduces the number of satellite links needed, as each cluster requires a minimum of only one satellite link. The various possible mesh network setups using the current geographical locations of the GPS station in the SuGAR array will also be presented.

In this study, each GPS station can be equipped with one or more long-range radios such as the FreeWave FGR-115RE. These radios specify a maximum range of over 90 km and can be used to form peer-to-peer wireless mesh networks between GPS stations. Assuming the maximum range of 90 km, the absence of relay stations or repeaters and the geographical locations of the 24 GPS stations, Fig.2 shows the network topology of GPS stations that will be formed using the FreeWave radios. It will contain one cluster with eight nodes, one

cluster with three nodes, two clusters with two nodes, and nine clusters with one node. Assuming that only one satellite uplink is required for each cluster, 13 satellite links will have to be maintained.

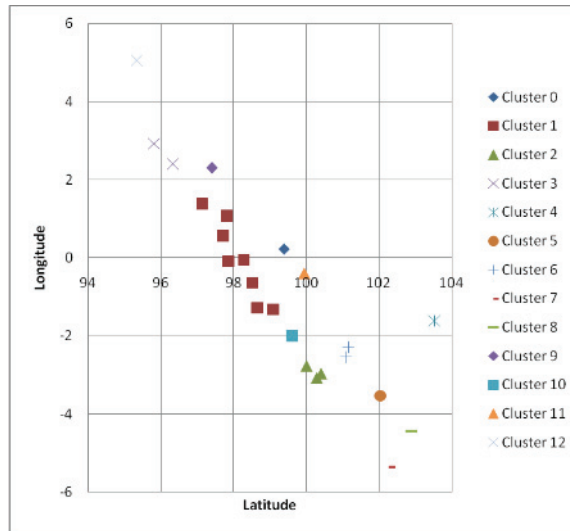


Fig. 2. Clusters of GPS station using 90 kilometer radio range

The range of the radio can be extended through the use of relay stations or repeaters. Thus, using the geographical locations of the 24 GPS stations, the minimum number of uplinks required and cluster size across various radio ranges can be determined. Fig. 3 shows the number of uplinks required for the various ranges. From the figure, it can be seen that given a maximum radio range of 20 km, only two GPS stations can be linked together and all other GPS stations were out of range from each other. Therefore, 23 satellite uplinks were required in this case. However, given a maximum radio range of 250 km, all GPS stations were grouped into one cluster using only one uplink.

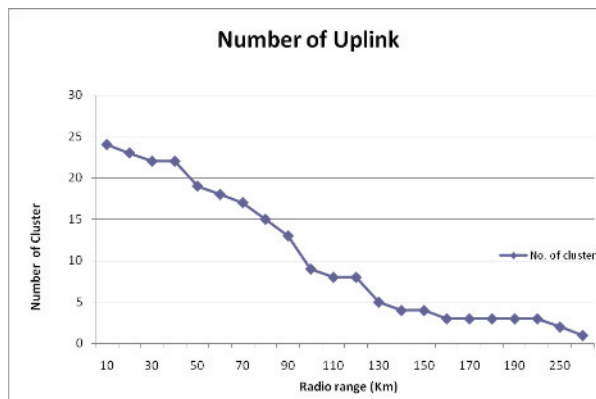


Fig. 3. Number of satellite uplinks required across various radio ranges

Fig. 4 provides the graph showing the average and the maximum number of GPS stations in a cluster across a radio range from 10 km to 250 km. As the number of GPS stations in a cluster increases, the data aggregated at the cluster-head will also increase in size. This will lead to better compression ratio at the cluster-heads and this phenomenon will be presented in more detail in the later part of this section.

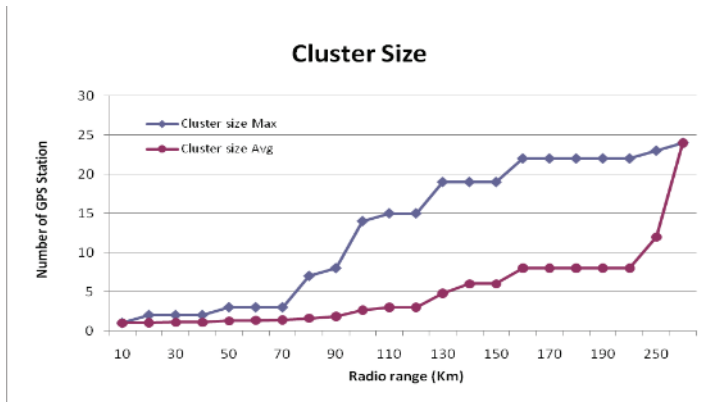


Fig. 4. Cluster sizes characteristics based on the various radio ranges

4.2 Collaborative compression of data

Cluster-based compression at the cluster-heads will be introduced where each cluster-head will compress the observation data from all GPS stations within the cluster using the LZMA (Ziv & Lempel, 1977) algorithm prior to transmission via the satellite link. Compared to the existing SuGAR deployment where each GPS station transmits the observation data independently, the use of mesh networking allows larger datasets to be formed through the aggregation of observation data from each GPS station within the cluster. Given that the compression ratio generally increases in proportion to the size of the dataset to be compressed, the number of bytes transmitted via the satellite will be significantly reduced.

Currently, the SuGAR sends collected data daily through dedicated satellite links from each GPS station. For this analysis, the GPS measurements will be converted locally to CRINEX format at each GPS station. Fig. 5 shows the total number of data bytes transmitted via all the satellite links using three different setups as follows:

- **Setup 1:** For the first setup, CRINEX data was uploaded via dedicated satellite links from each GPS stations without further compression.
- **Setup 2:** For the second setup, the CRINEX data was compressed using the LZMA algorithm prior to transmitting via dedicated satellite links at each GPS station.
- **Setup 3:** For the third and final setup, clusters of GPS stations were formed using long range radios with various maximum transmission ranges. In each cluster, one GPS station will be designated as the cluster-head and all other stations will forward their CRINEX data to the cluster-head. The cluster-head will perform further compression using LZMA algorithm on the aggregated data as a whole prior to transmitting the compressed data to the data server via a satellite link.

From Fig. 5, it can be seen that for Setup 2, the total number of bytes transmitted via all the satellite links over a 61 days period were reduced by about 67% when compared to Setup 1.

This demonstrates the effectiveness of the LZMA compression algorithm. Further reduction was demonstrated by the use of the cluster-based approach in Setup 3. In this setup, as a larger dataset was compressed, the compression ratios achieved by the LZMA algorithm at the cluster-head were more significant than in the case where compression was performed at individual GPS stations separately. Thus, this method reduced the total number of bytes transmitted by about 2% and 9% when compared to Setup 2 for a maximum radio range of 90 km and 250 km respectively.

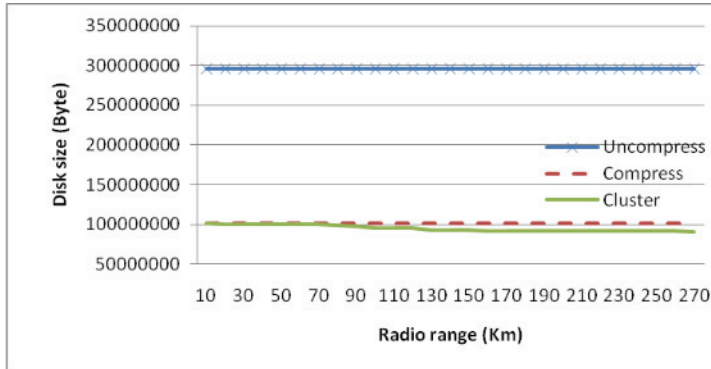


Fig. 5. Total size of transmitted data based on daily updates across two months (61 days) for various radio ranges

The analysis performed in Fig. 5 was based on daily updates from the GPS stations. However, more frequent updates might be useful for early warning systems and near real-time assessment of tectonic plate movements. Thus, further analysis was performed to evaluate the performance of the three setups across three different update intervals: daily, hourly and two minutely. Table 1 shows the comparison of Setup 1 (uncompressed data) and Setup 2 (un-clustered compressed data) with various update frequency. It can be seen from the results that as the update intervals get more regular, the performance of the LZMA algorithm suffers as smaller datasets were being compressed. For example, when daily updates were performed with the GPS station sampling once every 2 seconds, dataset consisting of a total of $(24\text{hrs} * 60\text{min} * 60\text{sec} / 2) = 43200$ measurements (epochs) was compressed whereas in the case where hourly updates were performed, each dataset consist of only $(60\text{min} * 60\text{sec} / 2) = 1800$ measurements (epochs). However, from the results, it can be seen that even when updates were performed every two minutes, the use of the LZMA compression in Setup 2 still enables less data to be transmitted via the satellite when compared to Setup 1.

Update Frequency	Total Transmitted Data		
	Uncompress	Compress	Percentage ^a
Daily	325,099,037 byte	112,188,360 byte	35%
Hourly	402,298,012 byte	158,994,711 byte	40%
2Minutely	2,245,193,111 byte	979,810,017 byte	44%

a. Percentage of compress data when compare with uncompress data

Table 1. Compare Uncompressed and Compressed Data

Fig. 6 shows the total transmitted data size in Setup 3 as a percentage to the total transmitted data size in Setup 2 across various radio ranges. From the results, it can be seen that the use of long range radios to form mesh networks and clusters in Setup 3 significantly reduces the amount of data to be transferred via the satellite links when compared to Setup 2. This reduction is more significant when the update frequency increases. This is due to the use of data aggregation within the cluster to enable larger datasets to be compressed. For example, when a maximum radio range of 250 km is used, data from all 24 GPS stations will be aggregated prior to compressing using the LZMA algorithm. Assuming hourly update intervals, each dataset consisting of $((60\text{min} * 60 \text{ sec} / 2) * 24 \text{ nodes}) = 43200$ measurements (epochs) was compressed in Setup 3 as compared to the 1800 measurements in Setup 2. Because of this, Setup 3 managed to reduce the total data transmission across the 61 days by about 70% when compared to Setup 2.

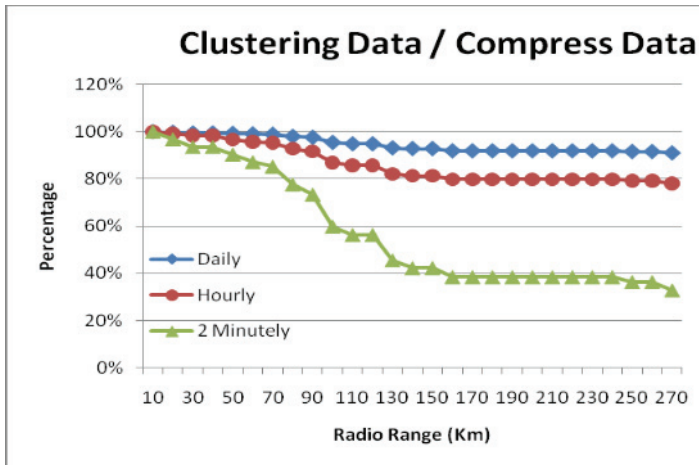


Fig. 6. Compare the improvement between compress observation data and use of clusters over different update intervals and radio range.

Update Frequency	Total Transmitted Data		
	Uncompress	Compress	Percentage ^b
Daily	322,554,780 byte	111,317,030 byte	35%
Hourly	341,813,991 byte	137,613,065 byte	40%
2 Minutely	710,381,007 byte	417,818,057 byte	59%

b. Percentage of compress data when compare with uncompress data without header

Table 2. Compare Uncompress and Compress Data without Header

To further reduce the size of the transmitted data, the observation headers sent with every update from the GPS stations were removed whenever possible. This significantly reduced the size of the uncompressed data in Setup 1 as shown in Table 2. Moderate reductions in Setup 2 were also observed when the observation headers were removed.

To conclude the evaluations, the use of Setup 3 (the use of wireless mesh networks) without observation headers was compared to Setup 2 (use of dedicated satellite links). The result of this comparison is shown in Fig. 7. From the figure, it can be seen that the use of mesh

networking, cluster-based compression and removal of the observation header significantly reduces the amount of data transmitted via the satellite links.

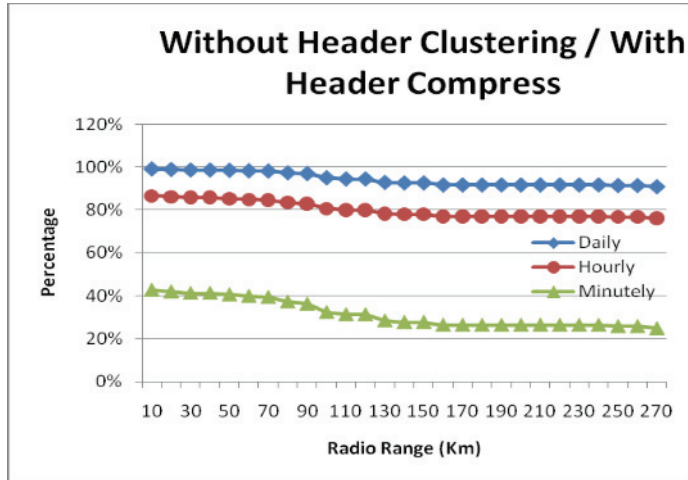


Fig. 7. Compare the improvement between compress observation data (with header) and use of cluster-based compression (without headers) over different update intervals and radio range.

4.3 Parallel and distributed in-situ processing for GPS corrections

In-situ parallel and distributed processing of GPS corrections can be made possible using mesh networking. The observation data from adjacent GPS stations can be grouped together and processed in a hierarchy fashion. Compared to the conventional method of sequential processing, the computational complexity and computation time of parallel and distributed GPS processing with various schemes decreases significantly. By sharing data within the mesh network, it is possible for in-network processing to be performed for GPS corrections using the embedded processing capability at each GPS station. This allows early-warning applications to be developed without the need for costly data transmission to a remote centralised server. The remaining of this section is organized as follow. Firstly, GPS measurement and parameters estimation process is briefly presented. Secondly, the computational complexity of parallel processing is evaluated using one layer and multiple layers approach. Finally, two empirical studies with various settings are studied.

Assuming that all receivers can receive signals from both frequencies L1 and L2, the ionosphere-free linear combination can be calculated. The distance between satellites and receivers are given by carrier phase and pseudo-range measurements. In phase measurement, at time t , the distance between receiver r and the satellite x models is derived as

$$L_{rxt} = \rho_{rxt} + b_{rxt} + z_{rt}m(\theta_{rxt}) + \omega_{rxt} + C_{rt} + c_{xt} + v_{rxt} \quad (1)$$

and the pseudo-range measurement is derived as

$$P_{rxt} = \rho_{rxt} + z_{rt}m(\theta_{rxt}) + C_{rt} + c_{xt} + \eta_{rxt} \quad (2)$$

in which, ρ_{rxt} is the true range, b_{rxt} is the phase bias or ambiguity, z_{rt} is the zenith troposphere delay, $m(\theta_{\text{rxt}})$ is the map function of elevation angle between transmitter and receiver. Receiver and transmitter correction are C_{rt} and c_{xt} respectively. The noise of the measurement is represented by v_{rxt} for phase and η_{rxt} for pseudo-range measurement.

Data is considered from R receiver and X transmitters spanning across Δ time with the data collection frequency σ . The median probability that a satellite signal is detected by a receiver above an elevation cutoff is given by $\Omega/4\pi$ (≈ 0.25 for a 15° cutoff). Thus, the number of measurement is given by

$$m = RX (\Omega/4\pi) (\Delta/\delta) d \quad (3)$$

in which d is the number of data types, typically including two types; ionosphere-free phase and pseudo-range. The number of parameters from those receivers and transmitters will be estimated and consist of receivers, transmitters and polar motion parameters. It is given by

$$n = aR + bX + c \quad (4)$$

The parameters related to the receiver include three Cartesian coordinates, tropospheric delay, receiver clock bias and phase bias parameter for each transmitter in the view of that receiver, so $a = 5 + X$. The transmitter parameters include epoch state position, velocity, two solar radiation parameters, Y bias parameter and clock bias, $b = 10$. Polar motion and rates are estimated in one day time given by $c = 5$.

The computation complexities of the parameter evaluation process using least square estimate method of n parameters with m measurement requires the number of arithmetic operations B in equation (5). This is also known as the computation burden. The detail analysis was presented in Zumberge, et al (1997).

$$B \propto n^2 m \quad (5)$$

One approach to reduce the computation complexity is to divide the data into groups and layers, which could then be processed in a parallel fashion. In addition, it makes use of common parameters and receivers between groups in the same layer. The detail of this processing approach will be presented in the next sub-sections.

4.3.1 Parallel GPS processing

In this part, parallel parameters estimation is studied with the objective of reducing the computation complexity and processing time when compared to the centralized processing method that is mentioned previously. It deals with estimating n unknown parameters of m measurements from R receivers and X transmitters. Moreover, receivers are divided into groups based on some criteria such as antenna type (Miyazaki, 1999), geography (Serpelloni et al., 2006), and/or the availability of data. Groups may share some common reference stations/receivers. One layer and multilayer parallel processing approach will be presented in the remaining of this section. All used notations are listed at the end of this chapter.

a. One layer parallelism

In one layer method, receivers are divided into J computation groups (Fig. 8) instead of estimating all parameters within one group. Suppose that the number of common parameters between all groups is kn and the remaining parameters equally divided for each group is $(1-k)n/J$. In addition, the number of common reference receivers between all

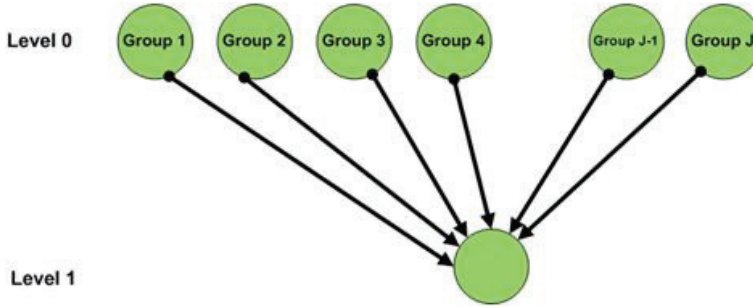


Fig. 8. One level parallel processing

groups is ζR . For simplicity, suppose the number of common measurement proportional to ζ is given by ζm and the remaining measurements are equally divided between groups, $(1 - \zeta)m/J$, for each group. The number of parameters and measurements at level zero for each group is thus derived as

$$n_{0,i} = \kappa n + \frac{(1 - \kappa)n}{J} \text{ and } m_{0,i} = \zeta m + \frac{(1 - \zeta)m}{J} \tag{6}$$

Arithmetic operations required are proportional to $n_{0,i}^2 m_{0,i}$, thus from equation (5)

$$B_{0,i} \propto \left(\frac{(1 + (J - 1)\kappa)n}{J} \right)^2 \frac{(1 + (J - 1)\zeta)m}{J} \tag{7}$$

in which $B_{0,i}$ is the number of arithmetic operations required at any group i ($1 \leq i \leq J$) at level zero. There are J groups in this level with the same number of arithmetic operations so the total number of operations is equal to J multiplied by the number of operation of one representative group $B_{0,1}$. Hence, the total number of arithmetic operations at level zero is equal to

$$B_0 = \sum_{i=1}^J B_{0,i} = J * B_{0,1} \tag{8}$$

Finally, the parameter estimation processing at level 1 is the refinement of J group at level zero. It includes n parameters and the number of measurement equaling to the total number of estimated parameter of J groups at level zero. Using equation (5), the computation burden is derived as

$$B_1 \propto n^2 \sum_{i=1}^J n_{0,i} = n^2 (1 + (J - 1)\kappa)n \tag{9}$$

Thus, the total number of operations B is equal to the sum of all computation burdens at level zero and level one as follows,

$$B = B_0 + B_1 \propto n^2 (1 + (J - 1)\kappa) \left(\frac{(1 + (J - 1)\kappa)(1 + (J - 1)\zeta)m}{J^2} + n \right) \tag{10}$$

The computation reduction percentage χ is equal to number of operations divide by the number of operation n^2m required for simultaneous parameter evaluation.

$$\chi = \frac{B}{n^2m} \propto (1 + (J - 1)\kappa) \left(\frac{(1 + (J - 1)\kappa)(1 + (J - 1)\zeta)}{J^2} + \frac{n}{m} \right) \tag{11}$$

The value of χ approaches unity when ζ and κ approaches 1 assuming n/m is small. Therefore, if all the parameters and receivers are common between groups, parallel processing is ineffective.

This method is applied for the Sumatra continuous GPS (cGPS) array (Tran & Wong, 2009) and the results are evaluated for two different configurations using the parameters $X = 24$, $\Omega/4\pi = 0.25$, $\Delta = 24h$, $\sigma = 2 \text{ min}$, $d = 2$, $a = 29$, $b = 10$, $c = 5$. For the first configuration, the number of receivers R equal to 40 which include 32 GPS stations of Sumatra cGPS array and 8 International GNSS Service (IGS) reference stations. In the second configuration, only 32 Sumatra cGPS stations were used without reference stations.

In the first configuration, we have ζ equal to the number of reference stations divide by the total number of stations, thus, $\zeta = 8/40 = 0.2$. The number of common parameters equal to the sum of the parameters of the common reference stations, the transmitter parameters and the polar motion. This can be calculated using equation (12), so $\kappa \approx 0.34$.

$$\kappa n = a\zeta R + bX + c \tag{12}$$

In the second configuration, the number of common reference stations, ζ , is equal to zero and so, using equation (12), $\kappa \approx 0.17$.

The computation reduction with respect to the different groups is presented in Fig. 9. In the case where reference stations were utilized, the maximum reduction reached 57% when receivers were divided into 5 groups. It decreases when the number of group increased due to the overheads of the reference station when using more groups. In the case where no reference stations were used, the maximum reduction reaches 91.6% when receivers were divided into 16 groups with 2 receivers per groups.

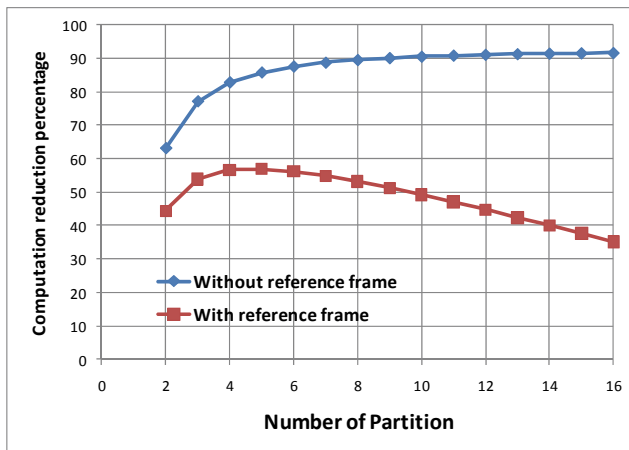


Fig. 9. Computation reduction for the Sumatra cGPS array using one level parallel processing

b. Multilayer parallelism

For generalization, the multilayer parallel is studied with L layer and each layer includes power of p groups. It denotes that there are p power of L groups at level zero and each group at level j (1≤j≤L) receives data from p groups at the adjacent predecessor level j-1. For instance, p equals to two in Fig. 10.

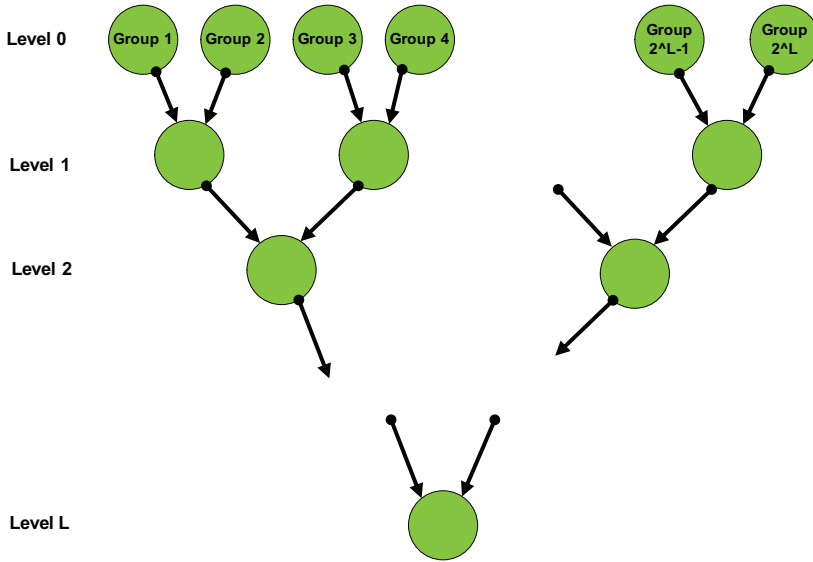


Fig. 10. Multilayer parallel processing with L layer with power of 2 groups. The processing tree will contain 2^L groups at level 0 and each group at level j (0<j≤L) is the combination of 2 node at level j - 1.

With the same assumption of common parameters and measurements with the one layer parallel method mentioned previously, the number of parameters is equal to the sum of the common parameters and private parameters of each group of receivers and number of measurements are equal to sum of the common measurements from common receivers and private measurements from the private receivers.

$$n_{0,i} = \kappa n + \frac{(1 - \kappa)n}{p^L} \text{ and } m_{0,i} = \zeta m + \frac{(1 - \zeta)m}{p^L} \tag{13}$$

Therefore, the number of arithmetic operations of group i at level zero is

$$B_{0,i} \propto n_{0,i}^2 m_{0,i} = \left(\kappa n + \frac{(1 - \kappa)n}{p^L} \right)^2 \left(\zeta m + \frac{(1 - \zeta)m}{p^L} \right) \tag{14}$$

So, the total computation burden for level zero which include p^L group equals to

$$B_0 = \sum_{i=1}^{p^L} B_{0,i} \tag{15}$$

Furthermore, the computation burden for each group i at level j ($1 \leq j \leq L$) is proportional to $n_{j,i}^2 m_{j,i}$, in which the number of parameter $n_{j,i}$ is equal to the sum of common parameters κn and the private parameters of p ancestor group at level $j-1$, each of which comprise of $\left((1 - \kappa)n * p^{j-1}\right) / p^L$ private parameters. Therefore,

$$n_{j,i} = \kappa n + \frac{(1 - \kappa)n}{p^L} p^j \quad (16)$$

In addition, the number of measurements at level j is equal to the summation of all estimated parameters of p ancestor at level $j-1$,

$$m_{j,i} = p\left(\kappa n + \frac{(1 - \kappa)n}{p^L} p^{j-1}\right) = p\kappa n + \frac{(1 - \kappa)n}{p^L} p^j \quad (17)$$

Therefore, the computation burden of each group i at level j equals to

$$B_{j,i} \propto \left(\kappa n + \frac{(1 - \kappa)n}{p^L} p^j\right)^2 \left(p\kappa n + \frac{(1 - \kappa)n}{p^L} p^j\right) \quad (18)$$

The total computation burden for level j which include p^{L-j} groups is then derived as

$$B_j = \sum_{i=1}^{p^{L-j}} B_{j,i} \propto \left(\kappa n + \frac{(1 - \kappa)n}{p^L} p^j\right)^2 \left(p\kappa n + \frac{(1 - \kappa)n}{p^L} p^j\right) p^{L-j} \quad (19)$$

The total computation burden of multiple parallel processing is equal to summation of computation of all level from level 0 to L as follows:

$$B = \sum_{j=1}^L B_j + B_0 \propto \sum_{j=1}^L \left(\kappa n + \frac{(1 - \kappa)n}{p^L} p^j\right)^2 \left(p\kappa n + \frac{(1 - \kappa)n}{p^L} p^j\right) p^{L-j} + \left(\frac{(1 - \kappa)n}{p^L} + \kappa n\right)^2 (1 + (p^L - 1)\zeta)m \quad (20)$$

c. Computation time

Assuming that the computation time is the dominant latency between processing groups at adjacent layer, the processing time of parallel GPS processing, in the worst case, is calculated by the summation of the maximum computation time at each layer at the critical computation path. The critical path for one layer and multilayer parallel processes is given in Fig. 11 and Fig. 12 respectively.

The computation time C is equal to number of arithmetic operation multiply by c , the computation time for each arithmetic operation. The equation for one layer and multilayer are therefore derived as follow:

$$C_{onelayer} = \left[n^2 (1 + (J - 1)\kappa)n + \left(\frac{(1 + (J - 1)\kappa)n}{J} \right)^2 \frac{(1 + (J - 1)\zeta)m}{J} \right] * c \quad (21)$$

$$C_{multilayer} = \left(\sum_{j=1}^L \left(\kappa n + \frac{(1-\kappa)n}{p^L} p^j \right)^2 \left(p \kappa n + \frac{(1-\kappa)n}{p^L} p^j \right) + \left(\frac{(1-\kappa)n}{p^L} + \kappa n \right)^2 \left(1 + (p^L - 1) \zeta \right) \frac{m}{p^L} \right) * c \tag{21}$$

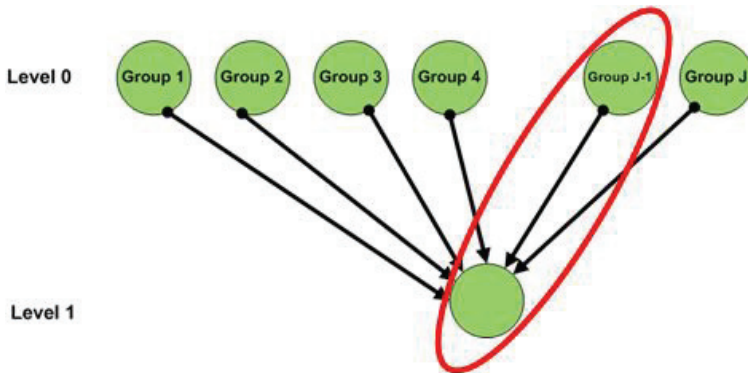


Fig. 11. One layer critical path

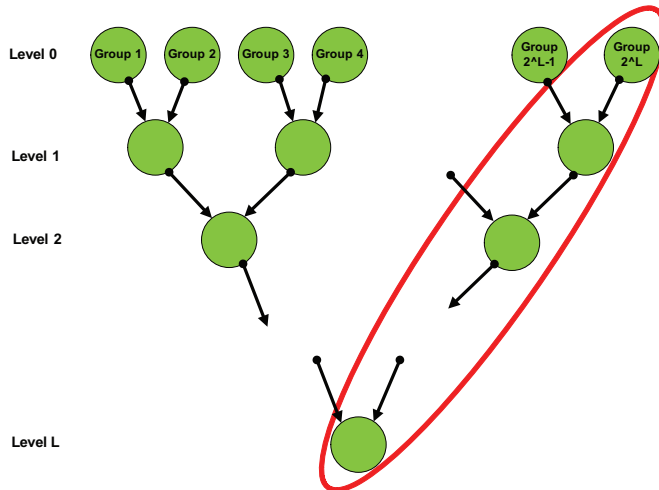


Fig. 12. Multilayer critical path

4.3.2 Empirical study

To compare the reduction in computation burden and computation time of one layer and multilayer parallel parameter estimation for GPS processing, two experimental setups were studied as following.

Experiment set 1: for the network parameter estimation, reference receivers were not included. This experiment compares the number of processing groups, computation reduction and computation time between three system settings with different number of GPS receivers. Three system settings are

- One layer,
- Multilayer with power of 2,
- Multilayer with power of 3

The results of experiment set 1 is shown from Fig. 13 to Fig. 15. From the results, it can be seen that when the number of receivers is equal to 16 or 48, the number of computation process for multilayer with power of 3 is smaller than other two settings. As a result, the computer reduction is lower than other settings and the computation burden is larger than multilayer with power of 2. With other number of receivers bigger than 48, the computation reduction is almost analogous for all settings. Parallel GPS processing significantly reduces the computation complexity, especially when the number of receivers is bigger than 32. Furthermore, multilayer processing drastically reduces the computation time by about 50% when compared with the one layer approach. In most of cases, the number of computation

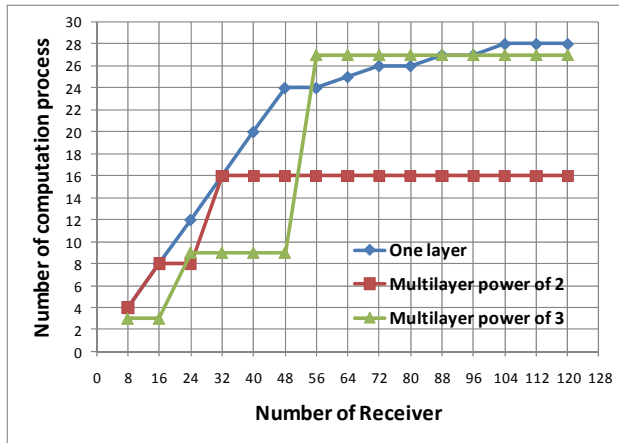


Fig. 13. Compare the number of computation processing groups with respect to number of receiver

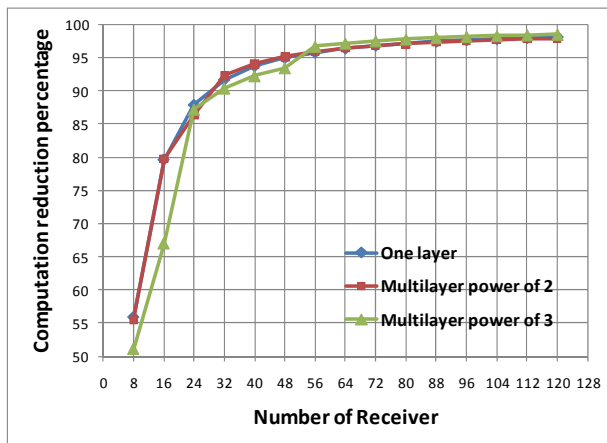


Fig. 14. Compare the computation reduction with respect to number of receivers

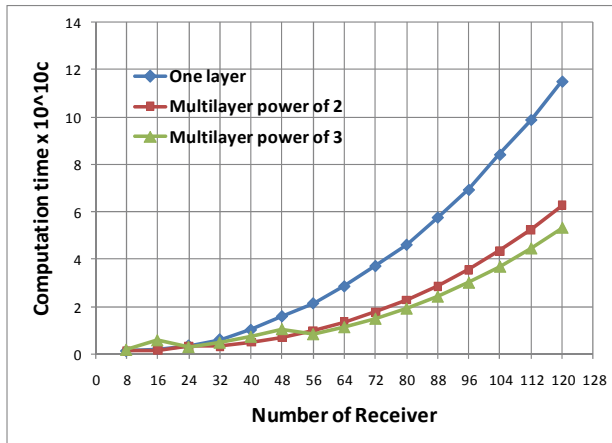


Fig. 15. Computation time comparison. The computation time is product of c , the computation time for one arithmetic operation

processes of multilevel methods is lower than one level method. As a result, multilevel is the best selection for in-network parameter estimation processing as demonstrated in this experiment.

Experiment set 2: global parameter estimate with 8 reference receivers (all group will share the same 8 reference receivers) using the same three comparative setting with the first experiment:

- One layer,
- Multilayer with power of 2,
- Multilayer with power of 3

The experiment results are shown from Fig. 16 to Fig. 18 (reference receivers are not included in the number receivers in the x-axis of the graph).

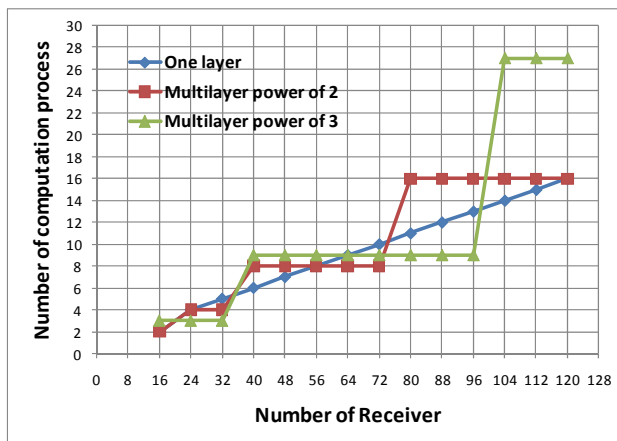


Fig. 16. Compare the number of computation processing groups with respect to number of receivers

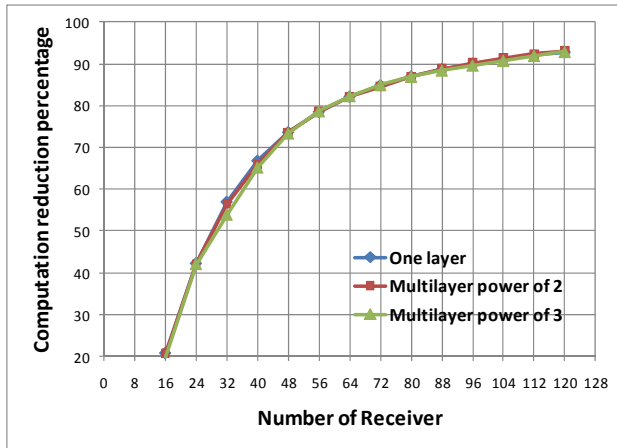


Fig. 17. Compare the computation reduction with respect to number of receivers

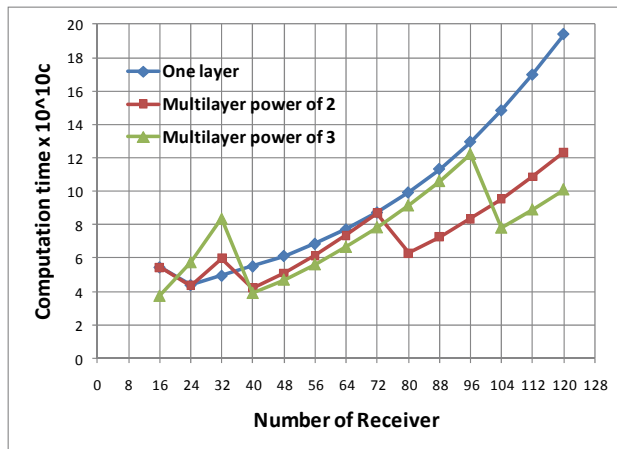


Fig. 18. Computation time comparison. The computation time is product of c, the computation time for one arithmetic operation

From the results, it can be seen that when the number of receivers is equal to 32 or 96, the number of computation processes using the multilayer approach with a power setting of 3 is smaller when compared to the other settings. The computation reduction is also larger than the other settings in the case of 32 receivers and larger than multilayer with a power of 2 in the case of 96 receivers. Thus, it can be seen that parallel GPS processing significantly reduces the computation complexity, especially when the number of receivers is bigger than 32 and steadily increases when the number of receivers increases. Furthermore, the multilayer processing approach slightly decreases the computation time, as in most of the cases, the number of computational operations performed by the multilevel methods is lower than the one level method.

5. Conclusion

A study using mesh networking for tectonic monitoring was presented. Mesh networks can be established between the GPS stations by means of long-range radios and data aggregation was performed to enable cluster-based compression. Using the actual data captured from the Sumatran cGPS array (SuGAR) in the evaluation and analysis, it was concluded that the proposed use of mesh networking not only reduces the number of costly satellite uplinks required, it also significantly reduces the total amount of data transferred through these links. Moreover, by making use of mesh networks between the GPS stations, parallel, distributed and hierarchical GPS processing methods can be made possible. By reducing the computation complexity, this proposed computational model allows the possible use of the spare computational power within the cGPS network such as from the routers and station controllers using the wireless mesh network connections between stations to transmit GPS data and perform collaborative GPS processing in a real-time fashion.

6. References

- Ammon, C. J., Ji, C., Thio, H.-K., Robinson, D., Ni, S., Hjorleifsdottir, V., et al. (2005). Rupture Process of the 2004 Sumatra-Andaman Earthquake. *Science*, 308(5725), 1133-1139. doi: 10.1126/science.1112260
- Gurtner, W., & Mader, G. (1990). Receiver Independent Exchange Format Version 2. *GPS Bulletin*, 3(3), 1-8.
- Hatanaka, Y. (1996, 17-20 September). A RINEX Compression Format and Tools. Paper presented at the Proceedings of ION GPS-96.
- Hudnut, K. W., Bock, Y., Galetzka, J. E., Webb, F. H., & W. H. Young. (2001). The Southern California Integrated GPS Network (SCIGN). 10th International Symposium on Crustal Deformation Measurement, 129-148.
- Konca, A. O., Avouac, J.-P., Sladen, A., Meltzner, A. J., Sieh, K., Fang, P., et al. (2008). Partial rupture of a locked patch of the Sumatra megathrust during the 2007 earthquake sequence. [10.1038/nature07572]. *Nature*, 456(7222), 631-635. doi: http://www.nature.com/nature/journal/v456/n7222/supinfo/nature07572_S1.html
- Lay, T., Kanamori, H., Ammon, C. J., Nettles, M., Ward, S. N., Aster, R. C., et al. (2005). The Great Sumatra-Andaman Earthquake of 26 December 2004. *Science*, 308(5725), 1127-1133. doi: 10.1126/science.1112250
- Miyazaki, S.-i. (1999). Construction of GSI's Nationwide GPS Array. Proceedings of the Joint Meeting of the U.S.-Japan Cooperative Program in Natural Resources Panel on Wind and Seismic Effects, 31, 518-528.
- Patterson, N., Gledhill, K., & Chadwick, M. (2007). New Zealand National Seismograph Network Report for the Federation of Digital Seismograph Networks Meeting, 2007. Perugia, Italy: 2007 FDSN Meeting.
- Segall, P., & Davis, J. L. (1997). GPS applications for geodynamics and earthquake studies. *Annual Review of Earth and Planetary Sciences*, 25, 301-336. doi: 10.1146/annurev.earth.25.1.301

- Serpelloni, E., Casula, G., Galvani, A., Anzidei, M., & Baldi, P. (2006). Data analysis of permanent GPS networks in Italy and surrounding regions: application of a distributed processing approach. [Article]. *Annals of Geophysics*, 49(4-5), 897-928.
- Sieh, K., Natawidjaja, D. H., Meltzner, A. J., Shen, C.-C., Cheng, H., Li, K.-S., et al. (2008). Earthquake Supercycles Inferred from Sea-Level Changes Recorded in the Corals of West Sumatra. *Science*, 322(5908), 1674-1678. doi: 10.1126/science.1163589
- Tran, H.-H., & Wong, K.-J. (2009). Mesh Networking for Seismic Monitoring - The Sumatran cGPS Array Case Study. Paper presented at the Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE.
- Yamagiwa, A., Hatanaka, Y., Yutsudo, T., & Miyahara, B. (2006). Real-time capability of GEONET system and its application to crust monitoring. *Bulletin of the Geographical Survey Institute*, 53.
- Ziv, J., & Lempel, A. (1977). A Universal Algorithm for Sequential Data Compression. *IEEE Transactions on Information Theory*, 23(3), 337 - 343.
- Zumberge, J., Heflin, M., Jefferson, D., Watkins, M., & Webb, F. (1997). Precise point positioning for the efficient and robust analysis of GPS data from large networks. *J. Geophys. Res.*, 102(B3), 5005-5017.

Notations

R	number of receiver (GPS station)
X	number of transmitter (satellite)
n	total number of parameter have to estimate
m	total number of measurement
κ	share parameters percentage between groups
ζ	share measurement percentage between groups
B	computation burden
J	number of computation group
L	number of processing level
p	in multiple level processing method, group at level i receive data from p group at level i-1
$n_{j,i}$	number of parameter at level j and group i have to estimate
$m_{j,i}$	number of measurement at level j and group i
$B_{j,i}$	computation burden at group i of level j
B_j	total computation burden at level j