ADVANCES IN ELECTRONICS AND TELECOMMUNICATIONS is a peer-reviewed journal published at Poznań University of Technology, Faculty of Electronics and Telecommunications. It publishes scientific papers addressing crucial issues in the area of contemporary electronics and telecommunications. Detailed information about the journal can be found at: www.advances.et.put.poznan.pl.

Journal Editorial Board

KRZYSZTOF WESOŁOWSKI, Editor-in-Chief
Poznan University of Technology
Piotrowo 3A, 60-965 Poznań, Poland
krzysztof.wesolowski@et.put.poznan.pl

ANNA PAWLACZYK, Secretary
Poznan University of Technology
Piotrowo 3A, 60-965 Poznań, Poland
anna.pawlaczyk@et.put.poznan.pl

ADRIAN LANGOWSKI, Technical Editor
Poznan University of Technology
Piotrowo 3A, 60-965 Poznań, Poland
adrian.langowski@et.put.poznan.pl

Wojciech Bandurski
Poznan University of Technology

Hanna Bogucka
Poznan University of Technology

Andrzej Dobrogowski
Poznan University of Technology

Anna Pawlak
Poznan University of Technology

Marek Domanski
Poznan University of Technology

Wojciech Kabaciński
Poznan University of Technology

Maciej Stasiak
Poznan University of Technology

Ryszard Stasiński
Poznan University of Technology

Paweł Szulakiewicz
Poznan University of Technology

Advisory Board

Flavio Canavero
Politecnico di Torino
Italy

Tadeusz Czachórski
Polish Academy of Science
Institute of Théoretical and Applied Informatics
Gliwice, Poland

Pierre Duhamel
CNRS - Supélec
France

Lajos Hanzo
University of Southampton
UK

Michael Logothetis
University of Patras
Greece

Józef Modelski
Warsaw University of Technology
Poland

Maciej Ogórzelew
AGH Technical University
Jagiellonian University
Cracow, Poland

John G. Proakis
University of California
San Diego, USA

Ralf Schäfer
Fraunhofer Heinrich-Hertz-Institut
Berlin, Germany
Note from the Editor-in-Chief
Krzysztof Wesolowski .......................................................... 1

Recent Advances and Future Trends in Wireless Communications
Hanna Bogucka ................................................................. 3

Spatial Multiplexing in Random Wireless Networks
K. Stamatiou, J. G. Proakis, and J. R. Zeidler ............................ 5

A Vision into Medium-Long Term Research in Wireless Communications
S. Benedetto and L. M. Correia ............................................ 13

Spectrum Occupancy in Realistic Scenarios and Duty Cycle Model for Cognitive Radio
M. López-Benítez and F. Casadevall ........................................ 26

A Review of OFDMA and Single-Carrier FDMA and Some Recent Results
C. Ciochina and H. Sari ....................................................... 35

Bayesian Foundations of Channel Estimation for Smart Radios
R. Couillet, A. Ancora, and M. Debbah .................................. 41

Adaptive Coded Continuous-Phase Modulations for Frequency-Division Multiuser Systems
A. Perotti, P. Remlein, and S. Benedetto ................................ 50

Iterative Joint Data Detection and Channel Estimation for DS-CDMA Systems in the Presence of Time-Varying Channels
E. Panayırıcı, H. Doğan, and H. A. Çırpan ................................. 59

High-Level Design Methodology for Ultra-Fast Software Defined Radio Prototyping on Heterogeneous Platforms
C. Moy and M. Raulet ......................................................... 67

Game Theory in Wireless Communications with an Application to Signal Synchronization
G. Bacci and M. Luise ......................................................... 86
Sociability Based Routing for Environmental Opportunistic Networks
R. Verdone and F. Fabbri ............................................................... 98

Trends in Adaptive Modulation and Coding
A. Zalonis, N. Miliou, I. Dagres, A. Polydoros, and H. Bogucka .................. 104

Considering Microelectronic Trends in Advanced Wireless System Design

Technology as a Need: Trends in the Evolving Information Society
G. E. Corazza, A. Vanelli-Coralli, and R. Pedone ........................................ 124
The journal you hold in your hands is the first issue of *Advances in Electronics and Telecommunications*, a journal newly founded by the Faculty of Electronics and Telecommunications, Poznań University of Technology. We hope that it will soon become a good and widely recognizable place for publication of valuable papers devoted to several topics in the area of contemporary electronics and telecommunications. The web page accompanying the journal (http://www.advances.et.put.poznan.pl/) presents a list of topics covering its spectrum of interest. Let us note that this list is certainly not complete and it will be gradually expanded along with development of new fields in electronics and telecommunications.

Founding a new journal is a serious challenge for the editors and supporting institutions. We have dared to undertake it. Fortunately, during our work on the first issue we have received a strong, essential support from NEWCOM++ – the Network of Excellence in Wireless Communications which groups the leading European universities in which research and education in wireless technologies are performed. This has allowed us to edit the first issue with the support of top European experts in wireless communications who supplied us with good and interesting papers. This is also a proof of the on-going scientific cooperation within the European Union. I would like to express my great thanks to all the authors who decided to submit their papers and I hope that it is not their last publication in our journal.

The editors of a well recognized scientific journal are usually supported by the Advisory Board, whose members give valuable hints to the editorial team and, to a certain degree, guarantee the quality of the papers and of the journal as a whole. At this place, I would like to express my deep gratitude to all the members of our Advisory Board for supporting our initiative and for trusting in our editorial project. We will do our best to achieve and keep the high scientific level of the journal.

*Advances in Electronics and Telecommunications* will be issued in a two-fold way, in the traditional paper form and in the electronic, open access form on the Internet. In the Internet era, scientific publications are much widely distributed over the network. In consequence, it has a positive impact on the popularity of the journal, citations of the papers published in it and, finally, it results in attracting better and better authors. We hope that our policy will find acceptance among the readers. Subsequent issues will have their leading topics. We follow excellent examples of this approach given by IEEE Journal on Special Areas in Communications and EURASIP Journal on Advanced Signal Processing. Despite leading themes of each issue, authors are welcome to submit papers in open call as well.

Last but not least, I would like to thank Prof. Hanna Bogucka, the editor of the first issue, whose enormous energy in putting the journal to run has had an inspiring influence on the whole editorial team. Many thanks to the journal web page administrator, Dr. Robert Kotrys who has professionally designed the Internet link of our journal with the readers, authors and reviewers. Dr. Adrian Langowski is another person who, thanks to his technical work, helped us to achieve a good technical standard of the journal.

I would like to wish our readers, authors, reviewers, respectable advisors and the whole editorial team a fruitful cooperation and frequent studies of the papers published in our journal. Let us hope that in our world in which every value has to be measured and compared with benchmarks it will quickly climb on the journal ranking lists, will be indexed in the leading data bases and will achieve a high impact factor.

Krzysztof Wesolowski

*Editor-in-Chief*
Recent Advances and Future Trends in Wireless Communications

Note from the Issue Editor

The opening issue of Advances in Electronics and Telecommunications collects original research contributions in the area of modern radio communications. It consists of papers authored by the researchers of the NEWCOM++ Network of Excellence in Wireless Communications – the leading European network in this thematic area – and presents their collective vision on recent advances and future trends in wireless communications.

NEWCOM++ project (contract no. 216715) is being implemented since 2008 under the European 7th Framework Programme, targeting the objective ICT-2007.1.1: The Network of the Future, mainly in the direction of “Ubiquitous network infrastructure and architectures”. The competence of this network includes topics from the wireless communication system physical and link layers, medium access control, radio-resource management, networks and their organization as well as topics concerning cross-layer design and optimization, related measurements and system validation. The project addresses also interdisciplinary issues of wireless communication system design, such as flexible platforms, wireless security, opportunistic and cognitive wireless access, spectrum and energy-efficiency.

Today, after the second year of NEWCOM++ operation and successful collaboration of its members, effective research-integration mechanisms have been put in motion, thus enabling consolidation of efforts towards the progress of wireless communications in Europe and world-wide. Some representative results of the NEWCOM++ partners scientific cooperation are presented in this issue.

The issue begins with an invited paper authored by Kostas Stamatiou, John G. Proakis (member of the Advances in Electronics and Telecommunications Advisory-Board) and James R. Zeidler: “Spatial Multiplexing in Random Wireless Networks”. The paper considers the performance of various spatial multiplexing techniques, and the number of streams maximizing the transmission capacity.

The following papers have been co-authored by the NEWCOM++ researchers, presenting their view on future trends in wireless communications and reporting recent results in the above-mentioned research areas.

The second invited paper in this issue, by Sergio Benedetto (the NEWCOM++ Scientific Director) and Luis M. Correia, presents “A Vision into Medium-Long Term Research in Wireless Communications”. It summarizes NEWCOM++ Vision Book, in which the medium- and long-term research tendencies are identified. In particular, the discussion concerns networks (heterogeneous and opportunistic, cognitive and cooperative, and the ultimate limits of wireless networks), bandwidth and energy-efficient radio access, multimedia signal processing for wireless delivery, the so-called hard technologies and Green Communications.

Then, we have a few papers reviewing current trends in the design of wireless communication system hardware platforms, physical and MAC layers: “A Review of OFDMA and Single-Carrier FDMA and Some Recent Results” by Cristina Ciochina and Hikmet Sari, “Trends in Adaptive Modulation and Coding” by Andreas Zalonis, Natalia Miliou, Ioannis Dagres, Andreas Polydoros and Hanna Bogucka, and “Considering Microelectronic Trends in Advanced Wireless System Design” by Dominique Noguet, Guido Masera, Venkatesh Ramakrishnan, Marc Belleville, Dominique Morche and Gerd Asheid.


The following papers report innovative research results in the area of software-defined, opportunistic and cognitive radio: “Spectrum Occupancy in Realistic Scenarios and Duty Cycle Model for Cognitive Radio” by Miguel López-Benitez and Fernando Casadevall, “High-Level Design Methodology for Ultra-Fast Software Defined Radio Prototyping on Heterogeneous Platforms” by Christophe Moy and Mickaël Raulet, and “Sociability Based Routing for Environmental Opportunistic Networks” by Roberto Verdone and Flavio Fabbri.

Another group of contributions present scientific results in the area of physical layer algorithmic design: “Bayesian Foundations of Channel Estimation for Smart Radios” by Romain Couillet, Andrea Ancora and

The paper “Game Theory in Wireless Communications with an Application to Signal Synchronisation”, by Giacomo Bacci and Marco Luise (the NEWCOM++ Administration and Management Director), considers the issue of game-theory-based techniques applied to resource allocation in wireless communication networks. Specifically, it focuses on the issue of allocating power resources to optimize the receiver performance in terms of CDMA spreading code acquisition.

This selection of papers by the world-class researchers is a good sample of NEWCOM++ scientific activities. I would like to thank all the authors for their interesting contributions to the opening issue of Advances in Electronics and Telecommunications.

Hanna Bogucka
Issue Editor
Spatial Multiplexing in Random Wireless Networks
Kostas Stamatiou, John G. Proakis, and James R. Zeidler

Abstract—We consider a network of transmitters, each with a receiver at a fixed distance, and locations drawn independently according to a homogeneous Poisson Point Process (PPP). The transmitters and the receivers are equipped with multiple antennas. Under a channel model that includes Rayleigh fading and path-loss, and an outage model for packet successes, we examine the performance of various spatial multiplexing techniques, namely zero-forcing (ZF), ZF with successive interference cancellation (ZF-SIC or VBLAST) and DBLAST. In each case, we determine the number of streams that maximizes the transmission capacity, defined as the maximum network throughput per unit area such that a constraint on the outage probability is satisfied. Numerical results showcase the benefit of DBLAST over ZF and VBLAST in terms of the transmission capacity. In all cases, the transmission capacity scales linearly in the number of antennas.

Index Terms—Poisson point process, spatial multiplexing, MIMO, outage probability, transmission capacity

I. INTRODUCTION

The study of random wireless networks has recently gathered a lot of momentum in the research community, e.g., see [1]–[4]. The main motivation behind this work is the use of tools from stochastic geometry in order to derive analytical results on how different physical, medium-access-control and network layer parameters affect the network performance. A central assumption is that the network consists of a Poisson Point Process (PPP) of transmitters, and each transmitter (TX) has a corresponding receiver (RX) at a given distance. The justification for the widespread use of this model is that it allows the analytical study of an ensemble of network topologies and captures the randomness of the node locations typical in networks without infrastructure such as ad hoc and sensor networks. The metric that quantifies the network performance is the transmission capacity, defined as the maximum spatial density of TX-RX links, multiplied by their rate, such that a certain constraint on the packet success probability is satisfied [2], [5]. Assuming that the channel - consisting of fading and interference - is constant during a packet slot an outage model may be employed for packet successes, i.e., a packet is successfully received if the mutual information of the channel realization is greater than the desired information rate. This translates to a requirement that the signal-to-interference-and-noise-ratio (SINR) is larger than a predetermined threshold.

In the context described above, and provided that the TX and the RX are equipped with multiple antennas, the objective of this paper is to shed light on the performance of certain spatial multiplexing techniques which require channel knowledge at the RX side1; zero-forcing (ZF), ZF with successive interference cancellation (ZF-SIC, also known as VBLAST [6], [7]) and DBLAST [7], [8]. The performance of these techniques is well understood for the fading and additive noise channel [7], but not so in a network environment where the interfering nodes are randomly placed.

A. Related work

The outage probability and transmission capacity for different spatial diversity techniques and single-antenna transmission was evaluated in [9]. One of the main results of this work was that, in the small outage probability regime, for maximal ratio combining (MRC) the transmission capacity scales as $N^{2/b}$, where $N$ is the number of RX antennas and $b$ is the path-loss exponent. The authors in [10] considered various multiple-input multiple-output (MIMO) techniques, including spatial multiplexing, as a component of a physical layer that employs frequency hopping and coding in combating interference. They arrived at similar scaling laws to [9] regarding the network throughput and the expected progress, albeit from a different analytical path. Multiple-antenna transmission with perfect channel knowledge at the transmitter was studied in [11] and the optimal number of spatial modes, in terms of maximizing the transmission capacity for a given density, was illustrated. More recently, multi-user techniques such as interference cancellation and space-division multiple-access have been considered in [12]–[14]. Specifically, in [13], it was shown that optimally selecting the number of cancelled nearby interferers results in a linear scaling of the transmission capacity with $N$, under single-antenna transmission.

B. Contributions

We first consider single-antenna (or single-stream) transmission and revisit the performance analysis of MRC in a Poisson field of interferers, deriving a compact analytical expression for the outage/success2 probability. It is shown that $N$ RX antennas provide an approximate gain of $N^{-2/b}$ in terms of spatial contention, i.e., the rate of increase of the outage probability as a function of the transmitter density, when the latter is zero. This result provides an alternative interpretation to the scaling law derived in [9].

1This assumption is made for the sake of simplicity of the communication protocol, as feedback to the TX is not required.

2The terms “outage” and “success” probability, since complementary, are used interchangeably throughout the paper.
We then turn our attention to multiple-antenna (or multistream) transmission and, employing our findings for the single-stream scenario, derive exact expressions and approximations to the outage probability for ZF, VBLAST and DBLAST. The optimal number of streams such that the transmission capacity of the network is maximized is determined for each of these techniques in the small outage probability regime. The trade-off lies in the fact that, introducing more streams can potentially boost the information rate of each link, but also increases the interference level in the network. For DBLAST specifically, it is shown that, for \( b \geq 4 \), it is optimal to use all transmit antennas, while, for \( b < 4 \), the number of streams must be judiciously chosen such that the optimal trade-off is achieved. Numerical results indicate that the benefit of DBLAST over ZF and VBLAST is significant in terms of the transmission capacity. For all spatial multiplexing techniques, provided that the number of streams is optimally chosen, the transmission capacity scales linearly in the number of antennas.

**C. Paper organization and notation**

The remainder of the paper is organized as follows. In Section II we describe in detail our system model. Section III is devoted to the analysis of the single-stream scenario and Section IV covers the extensions to the multiple-stream case. Our numerical results are outlined in Section V and Section VI concludes the paper.

We note the following regarding the notation: a zero-mean complex Gaussian random vector \( \mathbf{x} \), with covariance matrix \( \mathbf{Q} = \mathbb{E}[\mathbf{x}\mathbf{x}^H] \) is denoted as \( \mathbf{x} \sim \mathcal{CN}(0, \mathbf{Q}) \); the central chi-square distribution with parameter \( \frac{1}{2}l \) and \( l \in \mathbb{Z}^+ \), degrees of freedom is denoted as \( \chi^2(l) \); the \( l \times l \) identity and zero matrices are denoted as \( \mathbf{I}_l \), \( \mathbf{O}_l \), respectively; “\( \times \)” stands for “proportional to”, “\( \approx \)” stands for asymptotic equality and “\( \sim \)” denotes an approximate equality.

**II. SYSTEM MODEL**

The network consists of an infinite number of TXs, each with a corresponding RX at distance \( R \), and locations \( \{x_i\} \) that are drawn independently according to a homogeneous PPP \( \Pi = \{x_i\} \) of density \( \lambda \). Time is slotted and transmissions take place concurrently and in a synchronized manner during each slot. Due to the stationarity of the homogeneous PPP, the performance of any TX-RX link, i.e., “typical” link, may be studied. The network model, within a disc of finite radius around the typical RX, is depicted in Fig. 1.

The channel between each TX-RX pair consists of constant flat Rayleigh fading and path-loss according to the law \( r^{-b} \), with \( b > 2 \) (this requirement ensures that the interference power is finite [15]). Additive noise is disregarded, hence interference from concurrent transmissions is the only cause of errors in communication\(^4\). The power from each antenna is the same across all transmitters and, due to the absence of noise, may obtain an arbitrary value, e.g., unity. Generally, there is a different number of antennas at the TX and the RX; however, for convenience, we assume that \( N \) antennas are available at both the TX and the RX\(^3\).

Suppose that \( M \) antennas are employed for transmission, with \( M \leq N \). The received vector at the typical RX can be written as

\[
y = \mathbf{Hx} + \mathbf{w},
\]

where \( \mathbf{H} \) is the \( N \times M \) channel matrix between TX and RX, with i.i.d. elements \( \mathbf{H}_{mn} \sim \mathcal{CN}(0,1) \); \( \mathbf{x} \sim \mathcal{CN}(0, \mathbf{I}_M) \) is the \( M \times 1 \) symbol vector transmitted by TX; and \( \mathbf{w} \) is the interference term, modeled as \( \mathbf{w} \sim \mathcal{CN}(0, \mathbf{I}_N) \), where

\[
z = MR^b \sum_{i \in \Pi \setminus \{x_0\}} R_i^{-b} \] \hspace{1cm} (2)

is the total interference power over a given slot, per RX antenna; \( x_0 \) denotes the location of the typical TX and \( R_i \) is the distance of the interfering TX at location \( x_i \) from the typical RX\(^5\). It is known that \( z \) is an \( \alpha \)-stable random variable with stability exponent \( \alpha = 2/b \) [1], [4]. Its moment generating function (mgf) is given by

\[
\Phi_z(s) = \mathbb{E}[e^{-sz}] = e^{-cs^\alpha}, \quad s > 0,
\] \hspace{1cm} (3)

where the parameter \( c \) is defined as \( c = \frac{\lambda \pi R^2}{\Gamma(1 - \alpha)M^\alpha} \) and \( \Gamma(x), \quad x > 0 \) denotes the gamma function.

**III. SINGLE-ANTENNA TRANSMISSION (\( M = 1 \))**

Consider the transmission of a single stream, i.e., \( M = 1 \) and \( c = \lambda \pi R^2/\Gamma(1 - \alpha) \). Defining the desirable information

\(5\)This assumption is reasonable in an ad hoc network, where a node can be a TX or a RX at different times.

\(4\)We select to study an interference-limited scenario in order to focus on the effect of cochannel interference on the performance of the employed physical-layer techniques. The analysis can be generalized to include thermal noise.

\(3\)Note that, taking into account the fading from an interferer to a typical RX, the interference is generally correlated across the RX antennas. Assuming the interference is uncorrelated is a worst-case scenario, which simplifies the analysis.

Fig. 1. Network model. The black circles denote the transmitters and the green circles the corresponding receivers at distance \( R \). Solid/dashed lines denote useful/interfering signals.
rate as $R = \log(1 + \theta)$, where $\theta$ is an appropriate signal-to-interference-ratio (SIR) threshold, the success probability corresponding to (1) is given by [7]

$$
P_s = \Pr \left( \log \left( 1 + \frac{\eta}{z} \right) > R \right)
= \Pr \left( \frac{\eta}{z} > \theta \right),
$$

(4)

where $a = \|H\|^2$ is chi-square distributed with $2N$ degrees of freedom, i.e., $a \sim \chi^2_{2N}$. The respective outage probability is $P_o = 1 - P_s$.

A. Evaluation of $P_s$

The evaluation of $P_s$ requires the knowledge of the statistics of the SIR $\gamma = \eta/z$. In the following theorem, the complementary cumulative distribution (ccdf) of $\gamma$ is derived.

**Theorem 1** Let $\gamma = \eta/z$, where $\eta \sim \chi^2_{2N}$ and $z$ is an $\alpha$-stable random variable with mgf given by (3). The ccdf of $\gamma$, $\bar{F}_\gamma(x)$, is given by

$$
\bar{F}_\gamma(x) = e^{-cx^\alpha} + e^{-cx^\alpha} \sum_{k=1}^{N-1} \frac{1}{k!} \sum_{n=k}^{N-1} \frac{|\beta_n|}{n!} (cx^\alpha)^k,
$$

(5)

where

$$
\beta_n^k = \sum_{m=1}^{k} (-1)^{n-m} \binom{k}{m} (\alpha m)_n, \quad k = 1, \ldots, n
$$

and $(\alpha m)_n = \alpha m \ldots (\alpha m - n + 1)$ is the falling sequential product.

**Proof:** By the definition of $\bar{F}_\gamma(x)$, we have that

$$
\bar{F}_\gamma(x) = \Pr(a > xz) = \int_{0}^{+\infty} \bar{F}_a(xy)f_z(y)dy,
$$

(7)

where $\bar{F}_a(t)$ is the ccdf of $a$, given by

$$
\bar{F}_a(t) = e^{-t} - \sum_{n=0}^{N-1} \frac{t^n}{n!} = \frac{\Gamma(N, t)}{(N - 1)!}, \quad t > 0.
$$

(8)

Substituting (8) in (7), we obtain

$$
\bar{F}_\gamma(x) = \Phi_z(x) + \sum_{n=1}^{N-1} \frac{x^n}{n!} \int_{0}^{+\infty} y^n f_z(y)e^{-xy}dy.
$$

From the Laplace transform property

$$
f_z(y)y^n \overset{L}{\leftrightarrow} (-1)^n \frac{d^n \Phi_z(x)}{dx^n},
$$

(9)

it follows that

$$
\bar{F}_\gamma(x) = \Phi_z(x) + \sum_{n=1}^{N-1} \frac{x^n}{n!} (-1)^n \frac{d^n \Phi_z(x)}{dx^n}.
$$

(10)

Using identity 0.430.1, p.24, [16] for the $n^{th}$ derivative of a composite function, after some algebra, we obtain

$$
\frac{d^n \Phi_z(x)}{dx^n} = x^{-n}e^{-cx^\alpha} \sum_{k=1}^{N-1} \frac{\beta_n^k}{k!} (cx^\alpha)^k,
$$

(11)

where $\beta_n^k$ is defined in (6). Substituting (11) in (10) and regrouping terms results in

$$
\bar{F}_\gamma(x) = e^{-cx^\alpha} + e^{-cx^\alpha} \sum_{n=1}^{N-1} \frac{1}{n!} \sum_{k=1}^{N-1} \frac{1}{k!} \frac{|\beta_n|}{n!} (cx^\alpha)^k
$$

$$
= e^{-cx^\alpha} + e^{-cx^\alpha} \sum_{k=1}^{N-1} \frac{1}{k!} \sum_{n=k}^{N-1} \frac{1}{n!} \frac{|\beta_n|}{n!} (cx^\alpha)^k.
$$

(12)

In order to arrive at (5), we now need to show that $(-1)^n\beta_n^k \geq 0$. Once again, using the identity for the $n^{th}$ derivative of a composite function, $\beta_n^k$ can be written as the following derivative evaluated at $x = 1$.

$$
\beta_n^k = \frac{d^n (1 - x^\alpha)^k}{dx^n} \bigg|_{x=1}.
$$

(13)

From (13), the following iterative relation can be proved for $n \geq 2$

$$
\beta_n^k = \sum_{m_1=1}^{n} \left( \frac{n}{m_1} \right) \beta_{m_1}^{m_1} \beta_{n-m_1}^{m_1}.
$$

(14)

By successive application of (14), we obtain

$$
\frac{(-1)^n\beta_n^k}{n!} = \sum_{m_1=1}^{n} \sum_{m_2=1}^{n-m_1} \cdots \sum_{m_k=1}^{n-m_{k-1} - \cdots - m_1} (-1)^{m_1} \beta_{m_1}^{m_1} (-1)^{m_2} \beta_{m_2}^{m_2} \cdots (-1)^{m_k} \beta_{m_k}^{m_k},
$$

(15)

where $m_k = n - m_{k-1} - \cdots - m_1$. However, $(-1)^n\beta_n^k \geq 0$, since, by (6), $(-1)^n\beta_n^k = (-1)^{n+1} \alpha(\alpha - 1)\cdots(\alpha - n + 1)$ and $\alpha = 2/b < 1$. Therefore, $(-1)^n\beta_n^k \geq 0$ for $k = 1, \ldots, n$.

By the definition of $P_s$ in (4), we have that $P_s = \bar{F}_\gamma(\theta)$ or

$$
P_s = e^{-c\theta^\alpha} + e^{-c\theta^\alpha} \sum_{k=1}^{N-1} \frac{1}{k!} \sum_{n=k}^{N-1} \frac{|\beta_n|}{n!} (c\theta^\alpha)^k.
$$

(16)

We can see that $P_s$ is a product of the term $e^{-c\theta^\alpha}$ (the success probability for $N = 1$) and a polynomial in $c\theta^\alpha$ of degree $N - 1$ and non-negative coefficients. Clearly, increasing the number of antennas $N$, increases the success probability as more positive terms are added to the polynomial.

In order to obtain more insight into the effect of $N > 1$ on the success probability, we evaluate the spatial contention parameter

$$
\eta = -\frac{\partial P_s}{\partial \lambda} \bigg|_{\lambda=0},
$$

(17)

defined in [17] for single-antenna networks as the slope of the outage probability as a function of the density $\lambda$, at $\lambda = 0$. By its definition, the larger $\eta$ is, the sharper the increase of the outage probability as $\lambda$ increases. We have the following proposition.
**Proposition 1** In a network with single-antenna transmission \((M = 1)\), the spatial contention parameter \(\eta\) is
\[
\eta = \frac{c\theta^\alpha}{\lambda} \Gamma(N - \alpha) \frac{\Gamma(N)}{\Gamma(1 - \alpha)}.
\]

Proof: From the definition of \(\eta\) and (16), we have
\[
\eta = \frac{c\theta^\alpha}{\lambda} + \frac{c\theta^\alpha}{\lambda} \sum_{n=1}^{N-1} \frac{|\beta|}{n!} (-1)^n (\alpha)_n
\]
\[
= \frac{c\theta^\alpha}{\lambda} \sum_{n=0}^{N-1} \frac{(-1)^n (\alpha)_n}{n!} \quad | (\alpha)_0 \triangleq 1
\]
\[
= \frac{c\theta^\alpha}{\lambda} \sum_{n=0}^{N-1} \frac{(-1)^n (\alpha)_n}{n!} = \frac{c\theta^\alpha}{\lambda} \sum_{n=0}^{N-1} \frac{(-1)^n (\alpha)_n}{n!}
\]
\[
= \frac{c\theta^\alpha}{\lambda} \sum_{n=0}^{N-1} \frac{(-1)^n (\alpha)_n}{n!} = \frac{c\theta^\alpha}{\lambda} \sum_{n=0}^{N-1} \frac{(-1)^n (\alpha)_n}{n!}
\]
\[
= \frac{c\theta^\alpha}{\lambda} \sum_{n=0}^{N-1} \frac{(-1)^n (\alpha)_n}{n!}
\]
\[
\text{where (19) stems from the binomial identity for falling sequential products and (20) is the result of the successive application of the gamma function property} \Gamma(x + 1) = x\Gamma(x).
\]

For increasing values of \(N\), Stirling’s approximation yields
\[
\frac{\Gamma(N - \alpha)}{\Gamma(N)} \approx N^{-\alpha} \left(1 - \frac{\alpha}{N}\right)^{N-\alpha} \frac{1}{e^\alpha}.
\]

However, it is easy to verify that \(\lim_{N \to \infty} \left(1 - \frac{\alpha}{N}\right)^{N-\alpha} = e^{-\alpha}\), so
\[
\frac{\Gamma(N - \alpha)}{\Gamma(N)} \approx N^{-\alpha} e^{-\alpha}.
\]

As a result, for increasing \(N\), \(\eta \approx \pi R^2 \theta^\alpha N^{-\alpha}\). In other words, \(N\) provides an approximate gain of \(N^{-\alpha}\) in terms of spatial contention, or, equivalently, the antenna array at the RX effectively decreases the SIR threshold \(\theta\) by a factor \(N\). Note that (21) is quite accurate for relatively small values of \(N\), e.g., for \(N = 5\), the error is of the order of \(10\%\) for \(b = 4\), and \(5\%\) for \(b = 6\).

To conclude the analysis of the single-stream scenario, the following proposition provides an upper bound on \(P_s\) which is tight as \(\lambda \to 0\).

**Proposition 2** For single-antenna transmission, \(P_s\) is upper-bounded as
\[
P_s \leq \exp(-\eta \lambda) \triangleq P_u(\lambda, N)
\]
where the equality holds for \(N = 1\). Furthermore, for \(\lambda \to 0\), \(P_s \approx P_u(\lambda, N)\).

Proof: For \(N = 1\), \(P_s = e^{-c\theta^\alpha}\), so (22) holds as an equality. For \(N > 1\), we take the Taylor series expansion of \(P_u(\lambda, N)\) over \(\lambda\) and compare individual terms with (16). The reader may verify that it suffices to prove that
\[
A_k \triangleq \sum_{n=1}^{N-1} \frac{(-1)^n \beta}{n!} \leq \left( \sum_{n=1}^{N-1} \frac{(-1)^n \beta}{n!} \right)^k \B_k,
\]
with \(k = 1, \ldots, N - 1\). If \(k = 1\), (23) holds as an equality. For \(k > 1\), it holds that
\[
B_k = \sum_{n=1}^{N-1} \sum_{n=k+1}^{N-1} \beta \delta_1 \delta_2 \delta_3 \delta_4
\]
where, for convenience, we have defined \(\delta_1^n = \frac{(-1)^n \sigma^2}{n!}\). Moreover, by (15), we have that
\[
\frac{(-1)^n \beta}{n!} = \sum_{m_1=1}^{n-m_k} \sum_{m_2=1}^{n-m_k-1} \sum_{m_3=1}^{n-m_k-2} \sum_{m_4=1}^{n-m_k-3} \beta \delta_1 \delta_2 \delta_3 \delta_4
\]
where \(m_k = n - m_{k-1} - \cdots - m_1\). Substituting (25) and (24) in (23), we can see that (23) is a true statement. This is due to the fact that the summation that gives \(A_k\) is over a subset of the terms that are summed to give \(B_k\).

Finally, by the definition of \(\eta\), for \(\lambda \to 0\), \(P_s \approx 1 - \eta \lambda \approx P_u(\lambda, N)\).

As a result of Proposition 2, (22) can be used as an approximation to \(P_s\) in the small outage probability regime.

**B. Transmission capacity \((M = 1)\)**

We utilize the results of the previous subsection in evaluating the transmission capacity of the network, defined as the maximum network throughput per unit area, such that a constraint \(P_s = 1 - \epsilon\) is satisfied [2], [3], i.e.,
\[
TC_c = \lambda_c(1 - \epsilon)R_c
\]
where the maximum contention density \(\lambda_c\) is determined by the constraint \(P_s = 1 - \epsilon\). In the small outage probability regime, e.g., typically, \(\epsilon \leq 0.1\), we can invoke Proposition 2 to derive the following approximation to \(\lambda_c\)
\[
\frac{P_u(\lambda_c, N) \approx 1 - \epsilon}{\exp \left( - \frac{\Gamma(N - \alpha)}{\Gamma(N)} \lambda_c \pi R^2 \theta^\alpha \right) \approx 1 - \epsilon}
\]
\[
\lambda_c \approx \frac{\log(1 - \epsilon)}{\pi R^2 \theta^\alpha \frac{\Gamma(N)}{\Gamma(N - \alpha)}}.
\]

From (26) and (27), an approximation to the transmission capacity is thus
\[
TC_c \approx \log(1 + \theta) \frac{(\epsilon - 1) \log(1 - \epsilon)}{\pi R^2 \theta^\alpha} \frac{\Gamma(N)}{\Gamma(N - \alpha)}.
\]

As seen by (28) and (21), the transmission capacity of a single-stream system scales as \(\Theta(N^\alpha)\) in the number of RX antennas \(N\).
IV. MULTIPLE-ANTENNA TRANSMISSION ($M > 1$)

A. ZF

We now turn our attention to the case $M > 1$ - hence $c = \lambda R^2\Gamma(1-\alpha)M^\alpha$. Assume that each packet is transmitted over the same antenna during a slot with a rate $R = \log(1+\theta)$. If ZF is employed at the RX, the success probability for each stream, $P_{sf}^s$, is also given by (4), with the difference that $a$ is now chi-square distributed with $2(N - M + 1)$ degrees of freedom, as $M - 1$ degrees of freedom are sacrificed in order to cancel out inter-stream interference [7]. As a result, invoking (16),

$$P_{sf}^s = e^{-cd^\alpha} + e^{-cd^\alpha} \sum_{k=1}^{N-M} \frac{(cd^\alpha)^k}{k!} \sum_{u=0}^{N-M} \frac{\beta^u_k}{n!}. \quad (29)$$

From Proposition 2, we also have that $P_{sf}^s \leq P_u(\lambda, N-M+1)$.

The transmission capacity is now defined as

$$TC_{zf}^c = \lambda^\alpha \log(1+\theta) M R, \quad (30)$$

where the maximum contention density $\lambda^\alpha$ is determined by the constraint $P_u(\lambda^\alpha, N-M+1) \approx 1 - e$ for small $\epsilon$, or

$$\lambda^\alpha \approx \frac{\log(1-\epsilon)}{\pi R^2 \theta^\alpha} \frac{\Gamma(N-M+1)}{\Gamma(N-M+1-\alpha) M^\alpha}. \quad (31)$$

The transmission capacity is thus given by

$$TC_{zf}^c \approx \log(1+\theta) \frac{(1-\epsilon) \log(1-\epsilon)}{\pi R^2 \theta^\alpha} \frac{\Gamma(N-M+1) M^{1-\alpha}}{\Gamma(N-M+1-\alpha)}. \quad (32)$$

Due to (21), for large values of $N - M$, we have that

$$TC_{zf}^c \approx \log(1+\theta) \frac{(1-\epsilon) \log(1-\epsilon)}{\pi R^2 \theta^\alpha} \frac{\Gamma(N-M+1) M^{1-\alpha}}{(N-M+1)^{1-\alpha}}. \quad (33)$$

$TC_{zf}^c$ in (33) can be analytically optimized\(^7\) over the number of streams $M$ by allowing $M \in (0, +\infty)$ and setting the following derivative to zero

$$\frac{\partial}{\partial M} (N-M+1)^\alpha M^{1-\alpha} = 0. \quad (34)$$

After simple manipulations, we obtain $M_{zf}^c = (1-\alpha)(N+1)$. Since the constraints $M_{zf}^c \leq N$ and $M_{zf}^c \in Z^+$ must also be satisfied, the optimal number of streams is

$$M_{zf}^c = \min \left\{ \left\lfloor (1-\alpha)(N+1) \right\rfloor, N \right\} \quad (35)$$

where, with a slight abuse of notation $\left\lfloor (1-\alpha)(N+1) \right\rfloor$ denotes the closest integer number to $(1-\alpha)(N+1)$ that maximizes (33). Note that $(1-\alpha)(N+1) \leq N$ holds if and only if $\alpha \geq 1/(N+1)$ or $b \leq 2(N+1)$, which is valid for large $N$ as, typically, $b \leq 6$.

Setting $M = (1-\alpha)(N+1)$ in (33), we easily obtain that $TC_{zf}^c \propto \alpha^\alpha (1-\alpha)^{1-\alpha} (N+1)$, which implies that $TC_{zf}^c = \Theta(N)$. The linear scaling is the result of the appropriate choice of the number of streams such that the information rate per MIMO link is optimally traded off with the amount of interference introduced to the network. If, e.g., $M = 1$ or $M = N$, (32) reveals that $TC_{zf}^c = \Theta(N^\alpha)$ and $TC_{zf}^c = \Theta(N^{1-\alpha})$, respectively, i.e., the scaling is sublinear. This result is reminiscent of the one in [13]; the optimal contention density scales linearly in $N$ for large $N$, only when the number of cancelled interferers is a fraction of $N$. Interestingly, the optimal value of this fraction is also $1 - \alpha$.

B. ZF-SIC (VBLAST)

Suppose that the RX employs ZF-SIC (VBLAST), i.e., it cancels out each packet that has already been decoded. The spatial diversity order corresponding to the “worst” packet, i.e., the packet that is decoded first is $N - M + 1$. Given an outage constraint $\epsilon$ on this worst stream, the maximum contention density is also given by (31). Assuming that perfect interference cancellation takes place, i.e., there is no error propagation, the transmission capacity of VBLAST is given by

$$TC_{zf}^{vb} = \lambda^{\alpha} \log(1+\theta) \sum_{m=1}^{M} P_u(\lambda^{\alpha} M - m + 1), \quad (36)$$

as the spatial diversity order corresponding to each stream progressively increases as more streams are subtracted [7]. The summation term in (36) is the total throughput of all transmitted streams, with corresponding diversity orders $N - M + 1, \ldots, N$, ordered from the worst to the best.

Due to the complicated nature of (36), it is not possible to determine analytically the optimal number of streams. The optimization is performed numerically in Section V.

C. DBLAST

In Sections IV-A and IV-B, a packet is transmitted on the same antenna for the duration of a slot, thereby experiencing the same fading conditions across that slot. In DBLAST [7], [18], a packet is separated into segments which are transmitted across the antennas and time, such that each segment experiences different fading conditions. The segments are then detected at the RX by ZF-SIC and, once a packet is decoded, its contribution to the received signal is subtracted. It is known that DBLAST, in conjunction with appropriate coding, approaches\(^8\) the outage performance of the MIMO Rayleigh channel [7], i.e., for a total transmission rate $MR$, the packet outage probability is given by

$$P_{o}^{db} = P \left( \log \det \left( I_N + \frac{1}{z} H H^H \right) < M R \right). \quad (37)$$

In the system model we are investigating, this probability has to be evaluated over the distributions of $H$ and $z$. A way to approach this evaluation analytically is to recall that $P_{o}^{db}$ may be upper-bounded as [18]

$$P_{o}^{db} \leq P \left( \sum_{m=1}^{M} \log \left( 1 + \frac{a_m}{z} \right) < M R \right) = P \left( \prod_{m=1}^{M} \left( 1 + \frac{a_m}{z} \right)^{\frac{1}{M}} < 1 + \theta \right) \quad (38)$$

\(^7\)Note that we can also optimize over the SIR threshold $\theta$ as in [17].

\(^8\)In practice, DBLAST suffers from a rate loss due to initialization.
where \( \{ a_m \} \) are independent chi-square random variables with \( a_m \sim \chi^2_{2(N-m+1)} \). We observe that, for reasonable values of \( \theta \) (e.g., \( \theta = 5 - 20 \) dB), an outage roughly occurs when the interference power \( z \) obtains a large value. In this case, the geometric mean of \( \{ 1 + \frac{a_m}{2} \}^M_{m=1} \) is approximately equal to the arithmetic mean, thus

\[
P \left( \frac{M}{m=1} \left( 1 + \frac{a_m}{2} \right) < 1 + \theta \right) \approx P \left( \frac{1}{M} \sum_{m=1}^{M} a_m < M\theta \right).
\]

(The accuracy of this approximation is verified in Section V.) Moreover, \( \sum_{m=1}^{M} a_m \sim \chi^2_{2N_tot} \), where

\[
N_{tot} = \sum_{m=1}^{M} (N - m + 1) = \frac{2NM - M^2 + M}{2}.
\]

As a result, an approximation to the success probability for DBLAST, \( P_s^{db} \), can be obtained from (16) as follows

\[
P_s^{db} \approx e^{-c(M\theta)^{\alpha}} + e^{-c(M\theta)^{\alpha}} \sum_{k=1}^{N_{tot}^{-1}} \left( \frac{c(M\theta)^{\alpha}}{k!} \right) \sum_{n=k}^{N_{tot}^{-1}} \frac{|\beta_n^n|}{n!}.
\]

(41)

By Proposition 2, for \( \lambda \to 0 \), \( P_s^{db} \approx P_a(\lambda, N_{tot}) \), so, under a constraint \( P_{s_r}^{db} = 1 - \epsilon \), the optimal contention density for DBLAST is

\[
\lambda_{db}^{opt} \approx -\frac{\log(1 - \epsilon)}{\pi R^2 \theta^{\alpha}} \frac{\Gamma(N_{tot})}{\Gamma(N_{tot} - \alpha) M^{2\alpha}}
\]

and the respective transmission capacity is

\[
TC_{db}^{opt} \approx \log(1 + \theta) \frac{\epsilon - 1 - \log(1 - \epsilon)}{\pi R^2 \theta^{\alpha}} \frac{\Gamma(N_{tot}) M^{1-2\alpha}}{\Gamma(N_{tot} - \alpha)}.
\]

(43)

From (43) and (21), for large values of \( N_{tot} \), \( TC_{db}^{opt} \) can be approximated as

\[
TC_{db}^{opt} \approx \log(1 + \theta) \frac{\epsilon - 1 - \log(1 - \epsilon)}{\pi R^2 \theta^{\alpha}} 2^{-\alpha} (2N - M + 1)^{\alpha} M^{1-\alpha}.
\]

(44)

As in Section IV-A, letting \( M \in (0, +\infty) \) and setting the derivative of \( TC_{db}^{opt} \) with respect to \( M \) equal to zero, we obtain that \( M_{db}^{opt} = (1 - \alpha)(2N + 1) \). Under the constraints \( M \leq N \) and \( M \in \mathbb{Z}^+ \), the optimal number of streams for DBLAST is therefore

\[
M_{db}^{opt} = \min \left\{ [1 - \alpha(2N + 1)], N \right\}
\]

(45)

where, as in (35), with a slight abuse of notation \( [1 - \alpha(2N + 1)] \) denotes the closest integer number to \( (1 - \alpha)(2N + 1) \) that maximizes (44). Note that \( (1 - \alpha)(2N + 1) \leq N \) is only possible if \( \alpha \geq \frac{N + 1}{2N + 1} \). This implies that, if \( b \geq 4 \) (which is a typical value of \( b \) for ground propagation) transmission with all antennas maximizes the transmission capacity if the network is operated in the small outage probability regime.

We now investigate how \( TC_{db}^{opt} \) scales with \( N \). Letting \( M = M_{db}^{opt} \) in (44) (but omitting the operation \( \left\lfloor \cdot \right\rfloor \) for simplicity), we obtain that

\[
TC_{db}^{opt} \propto \begin{cases}
2^{-\alpha} \alpha^\alpha (1 - \alpha)^{1-\alpha} (2N + 1) & \alpha \geq \frac{N + 1}{2N + 1} \\
2^{-\alpha} (N + 1)^{\alpha} N^{1-\alpha} & \alpha < \frac{N + 1}{2N + 1}
\end{cases}
\]

(46)

(47)

As a result, in both cases, \( TC_{db}^{opt} = \Theta(N) \). Finally, comparing the optimized transmission capacity of DBLAST with that of ZF, we have that, for large \( N \) and \( \alpha \geq 1/2 \), \( TC_{db}^{opt} \approx 2^{1-\alpha} TC_{ZF}^{opt} \), while, for \( \alpha < 1/2 \), \( TC_{db}^{opt} \approx \frac{2^{-\alpha}}{\alpha^\alpha} TC_{ZF}^{opt} \). The gain in both cases is a direct consequence of the robustness of DBLAST with respect to the fading, as the information in each packet is coded and transmitted across all the antennas during a slot.

V. NUMERICAL RESULTS

In this section, we consider a network with default parameter values \( R = 20 \) m, \( b = 4 \), \( \theta = 6 \) dB. In Fig. 2, we plot the theoretical (eq. (16)) - and simulated success probability as a function of the PPP density when \( M = 1 \) and \( N = 4 \). The agreement between theory and simulation confirms the validity of the analysis in Section III. We also plot the upper bound to the success probability given by (22). As shown in Proposition 2, the bound becomes tight for values of the success probability greater than 0.8 (or, as the PPP density becomes progressively smaller).
In Fig. 3, the theoretical and simulated success probability of ZF and DBLAST are plotted vs. the density of the PPP for a system where \( M = 3 \) antennas are employed in each TX. The agreement between theory and simulation is once again very satisfactory, which, in the case of DBLAST, confirms the validity of the approximations in Section IV-C.

Fig. 4 shows the dependence of the transmission capacity on the number of transmitted streams for the three MIMO techniques considered in Section IV. The total number of antennas takes two values \( N = 4, 8 \), the propagation exponent is \( \theta = 6 \) and a constraint \( \epsilon = 0.1 \) is placed on the outage probability. The DBLAST transmission scheme results in higher transmission capacity compared to VBLAST or simple ZF. Moreover, the gain between VBLAST and simple ZF is marginal, which is attributed to the fact that, with VBLAST, the maximum contention density is still determined by the subchannel with the smallest diversity order.

In Fig. 5, the propagation exponent takes the value \( b = 4 \). As predicted in Section IV-C, activating all the TX antennas maximizes the transmission capacity for DBLAST. In the case of ZF, the optimal number of streams is dictated by (35). In the case of DBLAST, the optimal number of streams is 1 when \( N = 4 \), and 3 when \( N = 8 \). These numbers are in accordance with (45). In both cases is \( \Theta(\sqrt{N}) \). Setting \( M = N \) and \( b = 4 \) in (32) we can also see that \( TC^a \) is \( \frac{TC}{\Gamma(1/2)} \) which is confirmed by the plot.

In Fig. 5, the propagation exponent takes the value \( b = 4 \). As predicted in Section IV-C, activating all the TX antennas maximizes the transmission capacity for DBLAST. In the case of ZF, the optimal number of streams is dictated by (35). Overall, in both Fig. 4 and Fig. 5, the agreement between theory and simulation is satisfactory.

In Fig. 6 and Fig. 7, the transmission capacity and the respective - optimal - number of streams are plotted vs. \( N \) for DBLAST, ZF and MRC and an outage probability constraint \( \epsilon = 0.01 \). As predicted in Section IV, in the case of DBLAST and ZF, and optimally selected \( M \), the transmission capacity scales linearly in \( N \), while, in the case of MRC, it scales as \( N^\alpha = \sqrt{N} \). At \( N \geq 3 \), DBLAST provides a capacity gain of approximately 1.4 compared to ZF, which is in agreement with the \( \sqrt{2} \) gain predicted at the end of Section IV-C. Moreover, it is observed that, for \( N \leq 3 \), MRC and ZF result in approximately the same transmission capacity.
VI. CONCLUDING REMARKS

In this paper, we conducted a study of a multiple-antenna single-hop random network, where the locations of the transmitters are determined according to a homogeneous PPP. Assuming channel knowledge at the RX only and that interference from concurrent transmissions is regarded as noise, we first evaluated the outage/success probability for single-antenna transmission and MRC at the RX. We then used the results and insights from this analysis in order to evaluate the outage performance and the corresponding transmission capacity of MIMO techniques such as ZF, VBLAST and DBLAST. We determined the optimum number of streams such that the transmission capacity of the network is maximized in the small outage probability regime and quantified the capacity gain of DBLAST over ZF and VBLAST.

In conclusion, our results shed light on how MIMO techniques, which are well understood in the single-user context, affect the capacity of a random wireless network. At the heart of the analysis and the resulting design guidelines lies the PPP geometric model, which allows us to take into account the randomness in the locations of the interfering TXs in the statistics of the interference power seen at the typical RX.

REFERENCES


Kostas Stamatiou received his Diploma in Electrical and Computer Engineering from the National Technical University of Athens in 1995 and his M.Sc. and Ph.D. degrees in Electrical Engineering in 2004 and 2009, respectively, from the University of California San Diego. His Ph.D. thesis focused on the performance analysis of interference-mitigating techniques in the context of cellular and ad hoc networks. His current research interests lie in the area of random networks and stochastic geometry.


James R. Zeidler (M76-SM84-F94) is a Research Scientist/Senior Lecturer in the Department of Electrical Engineering, University of California, San Diego. He is a faculty member of the UCSD Center for Wireless Communications and the University of California Institute of Telecommunications and Information Technology. He has more than 200 technical publications and 13 patents for communication, signal processing, data compression techniques, and electronic devices. Dr. Zeidler received the Frederick Ellersick award from the IEEE Communications Society in 1995, the Navy Meritorious Civilian Service Award in 1991, and the Lauritsen-Bennett Award for Achievement in Science from the Space and Naval Warfare Systems Center in 2000. He was an Associate Editor of the IEEE Transactions on Signal Processing and a Member of the Technical Committee on Underwater Acoustic Signal Processing for the IEEE Signal Processing Society.
A Vision into Medium-Long Term Research in Wireless Communications

Sergio Benedetto and Luis M. Correia

Abstract—This paper summarises the 1st NEWCOM++ Vision Book. In it, the community of NEWCOM++ researchers, shaped under the common ground of a mainly academic network of excellence, has tried to distil their scientific wisdom into a number of areas characterised by the common denominator of wireless communications, by identifying the medium-long term research tendencies/problems, describing the tools to face them and providing a relatively large number of references for the interested reader. The identified areas and the researchers involved in their redaction reflect the intersection of the major topics in wireless communications with those that are deeply investigated in NEWCOM++. They are preceded by an original description of the main trends in user/society needs and the degree of fulfillment that ongoing and future wireless communications standards will more likely help achieving.


I. INTRODUCTION

VISION books are drafted in several contexts by international research fora (e.g., Wireless World Research Forum), technological platforms (e.g., European e-Mobility), specialised agencies, etc., and serve different purposes. By definition, they need to incorporate some “visionary” germs, and as such deal with medium-long term predictions. Some deal with the scientific/technological evolution, trying to predict the future of key enabling technologies by extrapolating today’s characteristics and up-to-date research trends. The authors are typically renowned scientists and industry leaders, and their reports are mainly aimed at influencing the policy makers at large, so as to gain resources in their broad area. Other focus on the societal needs of technology, and try to depict the future needs of prospective customers, so as to offer business hints to equipment manufacturers and service providers. A few envision a more comprehensive view of the future shaped by technological advances and its impact on the large scale societal behaviour, offering matter for sociologists’ analysis, and, to a more diffused scale, making human beings dream about the impossible to become possible and available.

In their attempt to predict future trends, vision books typically undertake one of two approaches: either they present easily foreseeable evolution of present technologies and applications, in which case of course the depicted scenario will become real in a short-medium time frame, or they launch themselves in the realm of disruptive technological breakthroughs and “killer” applications, where almost invariably they fail in their visionary scope. There are many examples of this incapacity to capture the great/little potential of scientific/technological advances: fibre optic technology (this year Nobel Prize recognised its pioneer, late as usual), the Internet protocol, Shannon theory, videotelephony, and cellular SMS (Short Message Service), among others.

The reason may be very simple: the great jumps ahead in human history are the result of revolutionary theories/technologies, rather than the evolutionary stretching (and as such, foreseeable) of existing ones. They amount to discontinuities, a kind of “nonlinear” progress that can be predicted by the fantasy of novelists like Jules Verne, rather than by realistic scientists well rooted in their time.

For this, and other reasons related to our particular and somewhat narrower perspective, this paper is not motivated by the ambition to present a future, comprehensive scenario of wireless communications in the decades to come. Rather, a community of researchers, shaped under the common ground of the academic network of excellence NEWCOM++, has tried to distil their scientific wisdom into a number of areas characterised by the common denominator of wireless communications, by identifying the medium-long term research tendencies/problems, describing the tools to face them, and providing a relatively large number of references for the interested reader, which resulted in the 1st NEWCOM++ Vision Book [1].

Sections II and III are devoted to networks, with emphasis on features that are likely to characterise their evolution: heterogeneity and opportunism (Section II), and cognition and cooperation (Section III). Section IV stems addresses information-theoretical ultimate limits through mathematical tools like game theory, stochastic geometry, and random graphs, and techniques to approach them, such as decentralised network control, coordination and competition strategies, network coding. In Section V the emphasis is on the maximisation of the throughput per unit of bandwidth through adaptive strategies in the radio access. Section VI refers to processing and reliably delivering of multimedia signals. Section VII is fully devoted to a deep insight into the present and future of “hard” technologies, i.e., silicon technologies, high computing

1A companion paper (G.E. Corazza, A. Vanelli-Coralli, R. Pedone, “Technology as a Need: Trends in the Information Society”), submitted to this same journal, contains an original description of the main trends in user/society needs and the degree of fulfilment that ongoing and future standards will more likely help achieving, so that part will not be treated here. This paper is also extracted from [1], and originated from G.E. Corazza (ed.), Report on Requirements and Constraints in Communication Technologies, ICT-NEWCOM++ Project, Deliverable DL.5.2, European Commission, Brussels, Belgium, Jan. 2010, http://www.newcom-project.eu.
power, great flexibility, and power efficient design. Section VIII deals with “green” communications, a concept drawing from the universally shared concern about sustainable development. Finally, Section IX draws the main conclusions of this paper, highlighting the key aspects.

II. HETEROGENEOUS AND OPPORTUNISTIC NETWORKS

Recent years have witnessed the evolution of a large plethora of wireless technologies with different characteristics, as a response of the operators’ and users’ needs in terms of an efficient and ubiquitous delivery of advanced multimedia services. The wireless segment of the network infrastructure has penetrated in our lives, and wireless connectivity has now reached a state where it is considered to be an indispensable service as electricity or water supply. Wireless data networks grow increasingly complex as a multiplicity of wireless information terminals with sophisticated capabilities get embedded in the infrastructure.

When looking at the horizon of the next decades, even more significant changes are expected, bringing the wireless world closer and closer to our daily life, which will definitely pose new challenges in the design of future wireless networks. In what follows, a vision is briefly described on some of the envisaged elements that will guide this evolution, and that will definitely impact of the design of the network layers.

A clear trend during the last decade has been a very significant increase in the user demand for wireless communications services, moving from classical voice service towards high bit rate demanding data services, including Internet accessibility everywhere and anytime. From a general point of view, the provision of wireless communication services requires from the operator perspective an appropriate network dimensioning and deployment based on the available technologies and in accordance with the expected traffic demand over a certain geographical area. The target should be to ensure that services are provided under specific constraints in terms of quality observed by the user, related both to accessibility (e.g., coverage area and reduced blocking/dropping probabilities) and to the specific service requirements (e.g., bit rate and delay), while at the same time ensuring that radio resources are used efficiently so that operator can maximise capacity.

The basic ingredients, allowing to successfully face the demand for higher traffic and bit rates, obey the following general principles of increasing:

- spectral efficiency, i.e., the number of bits/s that can be delivered per unit of spectrum,
- the number of base stations to provide a service in a given area, thus, reducing the coverage area of each one,
- the total available bandwidth, i.e., having a larger amount of radio spectrum to deploy the services.

As a consequence, the natural evolution that one can envisage for the next decade resides in the principles of heterogeneity due to the coexistence of multiple technologies with different capabilities and cell ranges, in the need to have smart and efficient strategies to cope with the high demand of broadband services, and also, on a longer term basis, in the introduction of flexibility in the way spectrum is managed.

The introduction of Self-Organising Networks (SON) functionalities, aiming to configure and optimise the network automatically, is seen as one of the promising areas for an operator to save operational expenditures. It is not difficult to envisage that in the next decades different networks will make use of it. Standardisation efforts are needed to define the necessary measurements, procedures and open interfaces to support better operability under multi vendor environment.

As a result of the above, SON has received a lot of attention in recent years in 3GPP [2] or in other groups such as the NGMN (Next Generation Mobile Networks) project [3], an initiative undertaken by a group of leading mobile operators to provide a vision of technology evolution beyond 3G for the competitive delivery of broadband wireless services to increase further end-customer benefits. It can be consequently envisioned that SON mechanisms will play relevant role in the mobile networks in the framework of the next decade.

Self-organisation functionalities should be able not only to reduce the manual effort involved in network management, but also to enhance the performance of the wireless network. SON functionalities include [4] self-:

- configuration: the process where newly deployed nodes are configured by automatic installation procedures to get the necessary basic configuration for system operation,
- planning: self-configuration comprising the processes where radio planning parameters are assigned to a newly deployed network node,
- optimisation: the process where different measurements taken by terminals and base stations are used to auto-tune the network targeting an optimal behaviour by changing different operational parameters,
- managing: the automation of Operation and Maintenance tasks and workflows, i.e., shifting them from human operators to the networks and their elements,
- healing: the process intending to automatically detect and to solve/mitigate problems, avoiding impact on users.

Opportunistic networking primarily stems from mobile ad hoc networking, that is, the research area that applies to scenarios where the complete absence of any supporting infrastructure is assumed. Opportunistic networks will spontaneously emerge when users carrying mobile devices, such as PDA (Personal Digital Assistant), laptops, and mobile phones, meet. According to the opportunistic networking paradigm, partitions, network disconnections and nodes mobility are more regarded as challenging chances rather than limiting factors or exceptions.

Mobility, for example, is considered as a way to connect disconnected network portions; partitions or network disconnections are not regarded as limitations since, in a longer temporal scale and by exploiting a store-and-forward approach, the information will finally flow from a source to the final destination. So, in opportunistic networks, delivering can be considered just a matter of time.

Opportunistic networking shares also many aspects with delay-tolerant networking, which is a paradigm exploiting occasional communication opportunities between moving devices. These occasional inter-contacts can be either scheduled or random, although conventional Delay-Tolerant Networks...
(DTNs) typically consider the communication opportunities as scheduled in time. Opportunistic networking and delay tolerant networking have been somehow distinguished so far in the literature, although a clear classification does not exist. However, as a common understanding, opportunistic networking could be regarded as a generalisation of delay-tolerant networking with no a-priori consideration of possible network disconnections or partitions.

Vehicular networks, enabling communication between different vehicles in a transportation system have been another topic that has received attention during the last years, driven mainly by the automotive industry, but also by public transport authorities, pursuing the increase of both safety and efficiency in transportation means. Although a lot of effort has been devoted, no solutions are already available to the mass markets allowing the automatic operation of cars and the communications among them. It is thus expected that the evolution of wireless communications in different aspects, such as ad-hoc and sensor networks, distributed systems, and combined operation of infrastructure and infrastructureless networks, can become an important step so that vehicular communications become a reality in the next decades.

From a more general perspective, the term Intelligent Transportation Systems (ITS) has been used referring to the inclusion of communication technologies to transport infrastructure and vehicles targeting a better efficiency and safety of road transport. Vehicular Networks, also known as Vehicular Ad-hoc NETWorks (VANETs), are one of the key elements of ITS enabling the one hand the inter-vehicle communication and on the other hand the communication of vehicles with roadside base station.

The evolution of Internet for the next decade has been coined under the term Future Internet, embracing new (r)evolutionary trends in terms of, e.g., security, connectivity and context-aware applications, in which again also wireless technologies will become an important and relevant element for the success of the different initiatives. In particular, one of the envisaged challenging goals for Future Internet is the possibility to interconnect not only people through computer machines (i.e., people connected everywhere and anytime) but also all type of unanimated objects in a communication network (i.e., connecting everything with everything), constituting what has been coined as the Internet of Things [5]. It presents a vision in which the use of electronic tags and sensors will serve to extend the communication and monitoring potential of the network of networks. This concept envisages a world in which billions of objects will report their location, identity, and history over wireless connections, so in order to make it true, this will require dramatic changes in systems, architectures, and communications.

Different wireless technologies and research disciplines can be embraced under the Internet of Things concept. Although Radio Frequency IDentification (RFID) and short-range wireless communications technologies have already laid the foundation for this concept, further research and development is needed to enable such a pervasive networking world, with the necessary levels of flexibility, adaptivity and security. Such technologies are needed as they provide a cost-effective way of object identification and tracking, which becomes crucial when trying to connect the envisaged huge amounts of devices. Also real-time localisation technologies and sensor/actuator networks will become relevant elements of this new vision, since they can be used to detect changes in the physical status of the different things and in their environment. The different things will be enabled with the necessary artificial intelligence mechanisms allowing them to interact with their environment, detecting changes and processing the information to even take appropriate reconfiguration decisions. Similarly, advances in nanotechnology enabling the manipulation of matter at the molecular level, so that smaller and smaller things will have the ability to connect and interact, will also serve to further accelerate these developments.

III. COGNITIVE AND COOPERATIVE NETWORKS

The traditional approach of dealing with spectrum management in wireless communications has been the definition of a licensed user granted with exclusive exploitation rights for a specific frequency. While it is relatively easy in this case to ensure that excessive interference does not occur, this approach is unlikely to achieve the objective to maximise the value of spectrum, and in fact recent spectrum measurements carried out worldwide have revealed a significant spectrum underutilisation, in spite of the fact that spectrum scarcity is claimed when trying to find bands where new systems can be allocated.

One of the current research trends in the spectrum management are the so-called Dynamic Spectrum Access Networks (DSANs), in which unlicensed radios, denoted in this context as Secondary Users (SUs) are allowed to operate in licensed bands provided that no harmful interference is caused to the licensees, denoted in this context as Primary Users (PU). The proposition of the TV band Notice of Proposed Rule Making (NPRM) [6], allowing this secondary operation in the TV broadcast bands if no interference is caused to TV receivers, was a first milestone in this direction. In this approach, SUs will require to properly detecting the existence of PU transmissions and should be able to adapt to the varying spectrum conditions, ensuring that the primary rights are preserved. Based on these developments it is reasonable to think that the trend towards DSANs has just started and that given the requirements for a more efficient spectrum usage, it can become one of the important revolutions in the wireless networks for the next decades, since it breaks the way spectrum has been traditionally managed.

Cognitive Radio (CR) is a paradigm for wireless communications in which either a network or a node change their transmission and/or reception parameters (signal format and bandwidth, frequency band etc.) to communicate efficiently, avoiding interference with licensed or unlicensed users. Most readers are already familiar with this notion, whilst not so many might have heard about the “sister” concept of Cognitive Positioning (CP) [7].

According to our general definition above, cognitive systems strive for optimum spectrum efficiency by allocating capacity as requested in different, possibly disjoint frequency bands.
Such approach is naturally enabled, by the adoption of flexible MultiCarrier (MC) technologies, in all of its flavours. Most current and forthcoming wideband standards for wireless communications are based on such multicarrier signalling technology, so that the signal allocated to each terminal is formed as the collection of multiple data symbols intentionally scattered across non-contiguous spectral chunks.

On the other hand, modern wireless networks more and more expect availability of location information about the wireless terminals, driven by requirements coming from applications, or just for better network resources allocation. Thus, signal-intrinsic capability for accurate localisation is a goal of 4th Generation (4G) networks. All signal processing techniques that can contribute to the provision of accurate location information are welcome in this respect. Such techniques can pair the ones that a cognitive terminal adopts to establish a reliable, high-capacity link.

The concept of cooperation actually emerged in the late sixties with the work of Van Der Meulen [8]. Interestingly, the capacity of this scheme is still an open problem today. More recently, the concept of relaying or cooperation has gained a lot of interest for several reasons:

- The potential offered by multi-antenna transmission and/or reception is now clearly established, and this technology, known under the generic name of MIMO, has found its way in a number of standards. However, the idea has emerged as whether different non-co-located entities could “form a coalition” to mimic in a distributed manner a multi-antenna system, thereby getting access to the benefits of MIMO in term of bit rate and/or diversity.
- A natural way to exploit this idea is to serve a user by means of two or more base stations (macro-diversity). Assuming a wired backhaul, the base stations know in the best case all the data and the channel state information of all the users in the cell cluster. An issue, to avoid very heavy signalling, is that of distributed solutions based on partial data and/or channel knowledge at the coordinating node.
- While the motivation behind the previous concept is mainly to avoid inter-cell interference, another motivation is associated with the issue of coverage and users that might be out of (good) reach by any base station. An emerging concept is that of a “popping up base station” with wireless backhaul, which would help one or many poor users. The issue of decoding strategy has to be considered.
- Moving to a totally different scenario, like wireless sensor networks, there is also a clear interest for cooperative solutions. The network is of the mesh type, rather than of the star one. The choice of the cooperating nodes may be based on several criteria or utility functions, incorporating not only rate and/or bit/packet error measures, but also penalty depending on the power used and/or the status of the battery of the possibly cooperating nodes.

An important concept that emerged recently and deserves further investigation is that of “network coding”, which shows promises, but also needs to properly encode and simultaneously relay the information of several users at the same time. Energy saving has to be considered as well. Transmission power is only one part of the global picture, and associated with any communicating nodes, there are additional power prices, like those associated with computation and security. It would be highly interesting to investigate the potential of relaying or cooperative communications at the light of a holistic analysis of power consumption.

While cognition and learning have received a considerable attention from various communities, the process of knowledge transfer, i.e., teaching, has received fairly little attention to date. A novel framework is introduced, “docitive radios” (“docere” means to teach in Latin), which relates to radios (or general entities) that teach other radios. These radios are not (only) supposed to teach the end-result (e.g., “the spectrum is occupied”) but rather elements of the methods to getting there. This concept mimics well our society-driven pupil-teacher paradigm, and is expected to yield significant benefits for cognitive, and thus more efficient, network operation. Important and unprecedented questions arise in this context, such as: “Which information ought to be taught?” and “What is the optimum ratio between docitive and cognitive radios?”

A high-level operational cycle of docitive radios extends the typical cognitive radio cycle [9] through the docitive teaching element, where each of these elements typically pertains to the following high-level issues:

- **Acquisition:** The acquisition of data is quintessential in obtaining sufficient information of the surrounding environment, which can be obtained by numerous methods.
- **(Intelligent) Decision:** The core of a cognitive radio is without doubt the intelligent decision engine, which learns and draws decisions based on the provided information from the acquisition unit.
- **Action:** An important aspect of the cognitive radio is to ensure that the intelligent decisions are actually carried out, which is typically handled by a suitably reconfigurable Software Defined Radio (SDR), and policy enforcement protocols, among others.
- **Docition:** An extension of the cognitive networking part is realised by means of an entity that facilitates knowledge dissemination and propagation. A significant and non-trivial extension to this docitive paradigm comprises dissemination of information which facilitates learning.

Docitive radios and networks emphasise on the teaching mechanisms and capabilities of cognitive networks, and are understood to be a general framework encompassing prior and emerging mechanisms in this domain. Whilst the exchange of end-results among cooperatively sensing nodes has been explored in the wireless communication domain and the joint learning via exchange of states has been known in the machine learning community, no viable framework is available to date which quantifies the gains of a docitive system operating in a wireless setting. Numerous problems remain, in the areas of Information Theory, Wireless Channel, PHY (Physical) and MAC (Medium Access Control) Layers, and Distributed Learning, among others.
IV. The Ultimate Limits of Wireless Networks

This section discusses the grand challenges associated with the holy grail of understanding and reaching the ultimate performance limits of wireless networks. Specifically, the next goal in the community is to realise the Future Internet.

The challenges associated with max-weight and backpressure type of network control strategies have been shown to be throughput optimal. Grand challenges include the decentralised light-weight implementation of these algorithms, as well as the issue of coping with flow-level and other type of dynamics. Understanding the deep structural properties of these policies and shaping them towards achieving the goals above in the presence of numerous resource and interference limitations would be an important step forward.

A fundamental premise in dynamic queues is that max-weight policies have been initially developed for systems consisting of a fixed set of queues with stationary, ergodic traffic processes. In real life scenarios, the collection of active queues dynamically varies, as sessions eventually end, while new sessions occasionally start. In many situations the assumption of a fixed set of queues is still a reasonable modelling assumption, since scheduling actions and packet-level queue dynamics tend to occur on a very fast time scale, on which the population of active sessions evolves only slowly. In other cases, however, sessions may be relatively short-lived, and the above time scale separation argument does not apply. The impact of flow-level dynamics over longer time scales is particularly relevant in assessing stability properties, as the notion of stability only has strict meaning over infinite time horizons.

Another key feature related to the dynamics and autonomic operation of future networks, is their decentralised architecture. In max-weight policies, the scheduling and resource allocation mechanisms are not amenable to distributed implementation because every node should be aware of all the queue lengths and the topology state variables in order to independently optimise its own strategy. Furthermore, even if the required information is available, the scheduling might become extremely complicated especially for large networks. Hence, it is evident that distributed implementations of optimum throughput techniques will be one of the main challenges of future networks. The maximum matching derivation is the issue that mostly challenges the distributed implementation of max-weight and backpressure routing and scheduling policies. Nevertheless, the problem of efficient and fully distributed implementation of max-weight policies is an important open problem the solution of which is highly related to achieving the holy grail of throughput optimal network control.

Game theoretic modelling is a means for understanding and predicting stable network operating points, namely points from which no node has an incentive to deviate. To this end, understanding interaction and convergence to equilibrium points is very important. Furthermore, the community should focus on the notions of competition and cooperation. For the former, non-cooperative interaction would be the best way to model the network, while for the latter, distributed optimisation approaches would be desirable.

In light of emerging future autonomous networks, game theory based analysis and modelling should be enhanced in order to satisfy long cultivated anticipation of research community. First, it is necessary to derive more detailed models that capture all aspects of the novel communication paradigms and therefore result to more stringent conclusions and realisable protocols. In this context, one should expect models which do not preclude the strategy space or the rules of nodes interaction. On the contrary, the possible actions of each player - node, the stages and the repetitions of interactions, the utility functions and other components of the game should be identified dynamically.

Additionally, in future large and decentralised networks each node will operate with limited information about other nodes and the network state. In this setting, it is interesting to study possible equilibrium points and their respective properties such as efficiency loss and computational complexity. In this context, understanding the repercussions of various types of learning algorithms and the way they impact convergence to equilibrium points should be properly addressed.

Moreover, the large number of nodes players yields for the employment of evolutionary game theory where nodes are grouped with respect to their behavioural profile. By observing the results of the repeated interactions, it is possible to identify the dominant strategies and therefore have significant insights to be used in the efficient protocol design.

Network coding has emerged as a revolutionary paradigm for information transfer in uni-cast or multi-cast traffic. Network coding turns out to be particularly effective and robust in environments of intermittent connectivity and continuous volatility. Network coding brings along a number of yet to be resolved challenges in resource efficient network operation, coding (traffic mixing) across the network, security and optimal resource utilisation.

Wireless communications, based on broadcast transmissions, represent the natural background of exploitation for network coding. Nevertheless, it should be taken into account that application of network coding in wireless and sensor networks presents implementation problems different from the wired network setting. The most salient aspect of wireless transmission is the use of broadcast via omnidirectional antennas. Broadcasting may beneficially provide collateral transmission to neighbouring nodes for free, but also causes interference. The use of network coding may significantly alter the nature of the interplay between advantages and drawbacks. Another key example is the communication/computation trade-off of using network coding within sensor networks. Because communication is relatively expensive compared to computation in sensor networks, network coding may offer substantial advantages in power consumption.

Moreover, network coding will have an emerging role within the context of cooperative networking. The aim of cooperation is that network communications become more reliable and efficient when nodes “support” each other to transmit data. The cooperation can be achieved enabling neighbouring nodes to share their own resources and their power with the hope that such a cooperative approach leads savings for both the overall network resources and power consumption. Network coding
must be seen as a form of user cooperation, where the nodes not only share their resources and their power but also their computation capabilities.

Stochastic geometry is a rich branch of applied probability that allows the study of physical phenomena characterised by highly localised events distributed randomly in a continuum, intrinsically related to the theory of point processes [10]. Although stochastic geometry has been used to characterise interference in wireless networks at early 70’s, its use for modelling communication networks is relatively new yet rapidly increasing. Stochastic geometry is considered as an enabler for addressing fundamental performance limits of massive and dense wireless networks through network information theory. The main facets of innovation here entail the understanding of modelling certain spatiotemporal node scenarios and protocol interactions with certain stochastic processes and probability distributions and assess their impact on analysing network performance.

Wireless Sensor Networks (WSNs) emerge as a particular class of wireless networks, expected to proliferate in envisioned sensory networks. Central to the reason of deployment of sensor networks lies the estimation of unknown parameters in the area. Key research issues include the sensor network control to fulfil estimation functionalities, realise opportunistic communication and networking under limited spectrum and energy budget. The notion of constrained cooperation is put forward as means for modelling a wide range of wireless network topologies, infrastructure-based or infrastructure-less. A number of challenges associated with sensor networks are:

- **Spectrum management:** One of the main issues of future WSNs is its operation in a crowded spectrum scenario where multiple WSNs, all of them application-specific, have to coexist.
- **Cross-layer designs (MAC/PHY):** A cross-layer design encompassing the MAC and PHY layers by adopting realistic multiple access schemes is essential; in these cases, the probability of collision turns out to be directly related to the metric of interest.
- **Opportunistic Communications:** In a multipoint-to-point network one can exploit the fact that different users act as different antennas obtaining the so-called Multi-User Diversity (MUD).
- **Network topologies:** Several topologies have to be further considered, like deploying a heterogeneous network, composed of sensors with different capabilities.

The massive body of theoretical work exemplifies the advantages of future wireless technologies, which will be based on far reaching processing abilities, harnessing thus the theoretical perspective into practical approaches.

V. BANDWIDTH AND ENERGY EFFICIENT RADIO ACCESS

Through the last two decades, the amount and diversity of services provided by wireless systems has been drastically transformed. Mobile (cellular) communication, for instance, is nowadays offering a wide variety of multimedia-data services, in contrast to the limited voice and very simple data services offered in the past. In Wireless Local Area Networks (WLANs), as another example, the ability to be on-line without needing a wired connection is not sufficient any more, and users expect to experience similar data speeds and QoS as with a wired connection. This has lead to a rapid increase in data rate requirements within the standards of new and upcoming wireless communication systems.

In order to increase data rates, a simple approach would be to increase the bandwidth allocated to a certain system. However, the proliferation of applications that use the air interface as the transmission medium limits the amount of available bandwidth in the radio-frequency spectrum, making it a very scarce and costly resource. Therefore, research over the last years has been focused towards improving the spectral efficiency of wireless communications, so that higher data rates can be achieved within a given bandwidth. For example, the development of multiple transmit antennas and their stream-multiplexing abilities, along with advanced receiver architectures, has been a critical step in achieving this goal.

This section aims to discuss the current theoretical and practical research topics that address the problem of bandwidth and energy-efficient radio access, as well as to provide a brief description of unsolved interesting problems for future research in this area. A discussion of coding/decoding techniques, as well as techniques that employ adaptive modulation and coding is performed, after which iterative techniques for receivers are addressed, concluding with cognitive spectrum usage (always for the purpose of advancing spectral utilisation) with a special focus on applied game theory for this goal.

The evaluation of the capacity bounds for various channel models and system scenarios is an active research topic since Shannons pioneer work [11]. The computation of such performance limits defines the framework for the design and development of optimal communication techniques. When the channel fading level is known at the transmitter, then the Shannon capacity is achieved by adapting the transmit power, data rate, and coding scheme relative to this known fading level. In such a fading environment, Adaptive Modulation and Coding (AMC) is a powerful class of techniques for improving the energy efficiency and increasing the data rate over the fading channel. Therefore, trellis and lattice codes designed for Additive White Gaussian Noise (AWGN) channels can be combined with adaptive modulation for fading channels, with the same approximate coding gains.

In general, coding techniques for the fading channel should be made by taking into account the distinctive features of the underlying model. There, one relevant question is how code design for the fading channel differs from that for the AWGN channel, assuming a flat, slowly fading channel plus AWGN. The quest for optimal coding schemes in such a fading environment leads to the development of new criteria for code design. If the channel model is not stationary, as it happens in a mobile-radio system, then a code designed for a fixed channel might perform poorly when the channel varies. Therefore, a code optimal for the AWGN channel may actually be suboptimum for a substantial fraction of time. In these conditions, antenna diversity with maximum-gain combining may prove indispensable: in fact, under fairly general conditions, a channel affected by fading can be turned
into an effectively AWGN channel by increasing the number of diversity branches. Another robust solution is based on bit interleaving, which yields a large diversity gain due to the choice of powerful convolutional codes coupled with bit interleaving and the use of a suitable bit metric.

The construction of optimal coding techniques is an area of intense scientific interest. Coding and decoding will play a central role in future wireless systems. Apart from improving the theory and practice of Low-Density Parity-Check (LDPC)/Turbo and related families of codes, extended research activity can be found for the improvement of coding and decoding techniques for convolutional codes too. For MIMO systems, space-time block codes can realise full diversity gain and decouple the vector-ML decoding problem into simpler scalar problems, which dramatically reduces receiver complexity. Space-time Trellis codes yield better performance than space-time block codes by achieving higher coding gain at the cost of increased receiver complexity. The area of MIMO communication theory is new and full of challenges. Some promising MIMO research areas are: MIMO in combination with OFDMA (Orthogonal Frequency Division Multiple Access) and CDMA (Code Division Multiple Access), new coding, modulation, and receive algorithms, combinations of space-time coding and spatial multiplexing, MIMO technology for cellular communications and adaptive modulation and link adaptation (AMC) in the context of MIMO.

The recent advances in iterative processing have allowed the design of iterative receivers achieving close-to-optimal performance under the design assumptions. However, many challenges are still left unsolved: the complexity issues of iterative receivers are one of the most important; due to the iterative nature of the receiver, it is clear that the number of operations to perform in these types of receivers will be significantly larger than in the case of a sequential receiver. Furthermore, due to the iteration with the channel decoder, some latency in the detection of the information bits is introduced, which might be unacceptable for certain kind of applications. In order to solve this, future research will have to focus on the development of low-complexity, high performance iterative solutions, which could be derived from the analytical design frameworks by constraining the set of solutions allowed. Also, research on the field of low-complexity decoding will be crucial for the applicability of iterative receivers in practical solutions, as it will significantly alleviate the computational burden of having to perform several successive decoding iterations of the same signal.

Most of the iterative solutions designed so far have only been derived under rather ideal conditions and only address questions such as channel estimation, detection and decoding. In the future, the design of iterative structures should also include considerations such as synchronisation, RF imperfections (e.g. phase noise) or estimation and mitigation of other system’s interference. While several analytical frameworks for the design of iterative solutions exist, each fulfilling different optimisation criteria, it is not clear which the best approach for a given problem is. Therefore, much of the research effort in the coming years will be put in the analysis and comparison between the different formal approaches, in order to achieve full understanding of their singularities, advantages and drawbacks. Information geometry might be an interesting tool to be used in this context.

As mentioned, flexible spectrum-access methods will be needed (along with others mentioned above, e.g., adaptivity, MIMO, and advanced coding) in order to increase the efficiency of the utilisation of the scarce spectrum resource. Besides the major objective of maximising spectral efficiency, another goal of modern radio network design is to rationalise the distribution of radio resources and the cost of their usage. This means that some rationalising mechanisms should be employed in order to balance efficiency versus fairness, high QoS versus high Quality of Experience (QoE) for both the licensed as well as unlicensed users of spectrum. Spectrally-efficient modulation techniques and associated waveforms need to be defined which would assure the opportunistic access to fragmented spectrum of dynamically-changing availability. Such techniques and waveform designs should also minimise (or at least limit) interference generated to other nodes and users.

Despite their advantages, the challenges posed by the OFDM-based or NOFDM-based CR still need to be met. Solutions should be found to specific problems such as high Peak-to-Average Power Ratio (PAPR), sensitivity to frequency offsets, and Inter-Carrier Interference (ICI), as well as in conjunction to other CR challenges, such as accurate and agile spectrum sensing, interference avoidance, cross-layer design and complexity, flexible RF front-end, etc. Multicarrier technologies will probably dominate wireless communication standards in the upcoming 10-15 years. As discussed above, they are also suitable for non-standardised, opportunistic and cognitive access to wireless networks and their spectrum resources. However, there are a number of issues that will be challenges for researchers, engineers and regulatory bodies. One is multi-band MC system design, in which the total bandwidth is divided into smaller parts. The challenges in designing a multiband MC transceiver include flexible broadband RF front-end with broadband antennas and their transmit/receive switches, complex RF circuit design, fast analogue-to-digital and digital-to-analogue converters of high dynamic range, wide-range frequency synthesisers, and so on. Other issues of the MC-based CR are related to the interference reduction between the users, sensing of the radio environment, synchronisation of non-contiguous MC signalling with missing pilots and possibly asynchronous transmissions of multiple users, and detection of the adopted transmission parameters or efficient signalling of these parameters.

VI. MULTIMEDIA SIGNAL PROCESSING FOR WIRELESS DELIVERY

Clearly, when one searches to build communication systems to mobiles or between mobile terminals with a very high capacity, one implicitly has in mind that some multimedia (high quality sound, still images, video, and more in the near future) has to be transmitted. Speech by itself would not justify such investments. Even when web surfing is considered, the files by themselves are not very demanding, only the
subparts containing images and multimedia are costly in terms of bit-rate. Therefore, it has been recognised that it is useful to tune the communication system globally, taking this fact into account. This approach leads naturally to what is known as cross layer optimisation and “Joint Source and Channel Coding/Decoding” (JSCC/D) approach.

The difficulty is that the layered architecture was designed in order to assign a specific task to each layer, the layers assuming that the previous one was successful in completing its task. Therefore, they are independent of each other, and, as a result, the delivery is independent of the type of signal. This universality of the network should not be put into question, even if some global optimisation is performed. The techniques cited above are relatively short term. While cross layer optimisation is already used, JSC Decoding should be usable soon, due to recent advances, and JSC Coding requires a further global optimisation, which may take some time.

However, the evolution of wireless multimedia will certainly go beyond the mere transmission of speech, sound, images, videos and html files, and such an expansion will require further scientific investment in a number of fields. Obviously, the evolution of wireless multimedia transmission and processing is driven by advances in devices, in the networks, and in the processing. Some of these enabling technologies, which are likely to have an impact in a foreseeable future, include 3D vision and augmented reality, with all the corresponding challenges in signal processing and devices, among others.

It is necessary to develop network coding algorithms tailored to video streaming applications. The algorithms should provide the following set of functionalities:

- Unequal error protection: The algorithms shall provide differentiated packet treatment, enabling a more reliable delivery of the most important video packets, through their identification and involvement into more coding operations, and the provision of smooth video quality.
- Multiple description coding: Decoding for video applications should provide incremental quality as more packets are received, e.g., by applying multiple description coding techniques.
- Fully distributed versus centralised network coding: Fully distributed algorithms are extremely simple, but suboptimal, while a centralised network code optimises the end-user probability of correct decoding based on full or partial knowledge of the packets they have already received; partially distributed algorithms should also be considered in order to limit the communication overhead.
- Low-complexity decoding: Although it has been shown that complexity is manageable, especially the wireless applications could benefit from simpler algorithms, therefore, it is foreseen to develop such algorithms using techniques borrowed from sparse random coding theory and digital fountain codes.

In many situations, when several mobile users are asking for a given media at the same time, they will not have the same link capacity, and some of them can have a very small one. Since the delay must remain small, retransmissions must remain limited (if not forbidden); moreover, each terminal may need a different signal quality. This is the motivation for the use of scalable signals (the bitstream contains the various qualities needed) and multiple descriptions (the more descriptions you get, the better the signal quality) on the source side. This has a number of consequences on the whole communication system, since all layers must be adapted to this point to multipoint situation.

Moreover, there is much to be gained in terms of bandwidth savings and performance improvement when considering a multicast situation: a number of registered users are requiring the same signal, may be with different qualities. Most known techniques for robust transmission of multimedia have to be adapted to this situation. Layered coding is known to be the theoretical solution from an information theory point of view, but a lot of work has to be undertaken in order to be able to build a full, practical system making optimal use of this property all along the protocol layers.

One should make the best use of any redundancy found in the bitstream, either left by the source encoder, or introduced by the protocol (headers, cyclic redundancy check, and checksums), the network (network coding), and the physical layer (channel coding and modulation). However, in this JSCD approach, even when protocol layers are involved, it is easily seen that one level of optimisation is not used: due to backward compatibility, the allocation of the redundancy along the whole system is not performed. This is the next step in this direction, which would result in a true joint source/protocol/network/channel optimisation.

Finally, the global trend can be seen to be a more global optimisation of the wireless network, including properties of the source into account. This should be done ideally without
losing the universality of the network (which should still be able to carry various applications and signals). The tendency towards a variety of application and service (more than plain transmission) can also be seen in the recent years. Therefore, it seems that on top of the integration of the various aspects: source, channel, protocol, network, the service should also be integrated (more than just the source). In this sense, one will then be able to have true multimedia available, rather than simple video transmission. Such applications could be augmented reality (see above), machine to machine communication, content retrieval, and many situations that are currently studied in the multimedia setting.

VII. ENABLING “HARD” TECHNOLOGIES

Wireless technology has been benefiting from the advances the silicon technology has offered in the 90s and 00s. The whole telecommunication mutation from the analogue domain to the digital realm has in fact been made possible by the miniaturisation of transistors, leading to higher density though with low power efficient Integrated Circuits (ICs). In turn, this move to digital communication has created the boom or the digital ICT at all layers, from broadband communications to multimedia services. With this in mind, it would not make sense to foresee what telecommunication will offer in the future without considering the trends in silicon technology research and industry.

The key factor behind the digital revolution has been the CMOS (Complementary MetalOxideSemiconductor) technology down-scaling following the so-called Moores Law, which coined in the early 70s that the number of transistors would double every year. Although this rule has been validated over the 40 last years by the silicon industry and by the International Technology Roadmap for Semiconductors (ITRS) [12], it is agreed that we are coming to a new era where this rule is no longer valid. Several reasons can be identified:

- the physics of silicon introduces side effects in deep submicron technology;
- predictability of transistors behaviour is getting less accurate, leading to lower yield or less optimal usage of silicon;
- power density is going to levels beyond what cooling can offer;
- static power increases, which makes the assumption that the overall power consumption decreases as transistors shrink no longer valid;
- investments needed for new deep submicron CMOS are being so huge that only less than an handful of application justifies it.

For all these reasons, it is likely that we are on the verge of significant changes in the silicon capability roadmap, which make the analysis of future trends useful. Indeed, it is foreseen that the roadmap has to move from a pure down-scaling to new functionalities and combined technology-system innovation in order to manage future power, variability and complexity issues. However, there is no accepted candidate today to replace CMOS devices in terms of the four essential metrics needed for successful applications: dimension (scalability), switching speed, energy consumption and throughput [13]. Moreover, when other metrics such as reliability, designability, and mixed-signal capability are added, the dominance of CMOS is even more obvious. It is then realistic to think that other micro- or nano-technologies should be seen in the future as an add-on to CMOS and not as a substitute for it. This transition between the “business as usual” era and the entry to the post 2015 period, where new alternative or complementary solutions need to be found, needs to be addressed.

Bearing in mind this disruptive future and rather than extending the technology evaluation proposed by the ITRS, technology analysis carried out in Europe by MEDEA experts [14] suggests considering 3 major paradigms:

- More Moore: corresponding to ultimate CMOS scaling. This will be essential to supply the massive computing power and communication capability needed for the realisation of applications at an affordable cost and a power efficiency exceeding 200 GOPS/Watt (Giga Operations per Second/Watt) for programmable and/or reconfigurable architectures [15].
- More than Moore: corresponding to the use of heterogeneous technologies, such as MEMSs (Micro Electro-Mechanical Systems) or MEOMSs (Micro Electro-Opto-Mechanical Systems). This approach intends to address parallel routes to classical CMOS by tackling applications for which CMOS is not optimal, which can be classified in three major categories: interfacing to the real world, enhancing electronics with non-pure electrical devices, and embedding power sources into electronics.
- Beyond CMOS: corresponding to nanotechnology alternatives to CMOS. This paradigm intends to identify technologies that could replace CMOS, either in a disruptive or evolutionary way after CMOS will reach its ultimate limits.

The context of coexistence and convergence is driving demand for flexible and future-proof hardware architectures offering substantial cost and power savings. Manufacturers have already started activities towards the provision of multi-mode handsets featuring the advancements of the recent radio access technologies (e.g., 3GPP LTE, WiMAX IEEE 802.16m, and DVB-T2/H). However a large gap is growing in the field of flexible radio between advances in communication algorithms, methods, system architectures on one side, and efficient implementation platforms. Despite the large amount of available results in the system level technologies related to multi-mode, multi-standard interoperability and smart use of available radio resources (e.g., AMC and cross layer optimisation), a very limited number of hardware solutions have been proposed to really support this flexibility and convergence by means of power efficient, low cost reconfigurable platforms. In most cases, chipset vendors offer different solutions for each combination of standards and applications to be supported.

To enable a single modem to service multiple different wireless systems, highly flexible solutions are needed. In practice, currently implemented flexible hardware modems are focused on the receiver segment placed between the RF (Radio Frequency) front end and the channel decoder. In this part of
the modem, several digital signal processing algorithms, such as equalisation, interference cancellation multipath correlation (rake receiver), synchronisation, quadrature amplitude mapping/demapping and FFT (Fast Fourier Transform) can be run on vector processors, which allow for GOPS rates resorting to very high level of parallelism. However, other functional components of a modern modem (such as channel decoding) are not efficiently supported by vector processors and alternatives to software programmable architectures are not considered solid solutions: cost of hardwired dedicated building blocks becomes rapidly unacceptable with the number of standards to be supported, while reconfigurable hardware, such as FPGAs (Field Programmable Gate Arrays), are too expensive in terms of silicon area and standby energy consumption (which is due to leakage current, proportional to area).

SDRs, with cognitive capabilities, are getting prominence as potential candidates to meet the future requirements of mobile devices. Compared to the pragmatic design approach for flexibility motioned above, the SDR approach aims at providing a comprehensive design framework encompassing platforms, architectures, software, methodology and design tools. The paradigm of SDR poses new challenges or makes current design challenges more stringent. The most relevant ones are:

- **Portability**, which can be defined as the inverse of porting effort, represents the ease with which one waveform can be moved to another hardware platform, requiring a platform independent waveform description.
- **Efficiency** with respect to area and energy is essential, in order to decrease the power/energy consumption and extend the battery life, requiring high efficiency in waveform implementation.
- **Interoperability** denotes the ability that a waveform implemented on two different hardware platforms interoperates with each other.
- **Loadability** illustrates the ease with which a waveform can be loaded, over-the-air, into a hardware platform, programmed, configured and run, which can be increased by well defined and known interfaces in waveform implementation.
- **Trade-offs** between flexibility and efficiency becomes challenging in the wake of their contradictory nature, making heterogeneous Multi-Processor System-on-Chips (MPSoCs) an inevitable candidate as the hardware platform for implementing a waveform.
- **Cross layer design and optimisation** techniques are getting popular, if not mandatory, in order to cope with the increasing need for spectrum and energy efficiency, leading to very tight dependencies, interactions between physical and MAC, higher layers that have cognition, requiring flexibility in implementation and algorithms.

The complexity of modern, flexible implementation structures would hardly be manageable and their development is a tedious and error-prone task. What is needed is a description method that can lead to a (semi-)automatic generation of a waveform implementation directly from the specification. Therefore, a methodology is required, that raises the abstraction level of receiver design to make it manageable.

The digital communication research on multi-standard radio has started based on the assumption of SDR, which extrapolated that RF stages of a radio would be transparent for the baseband processing either thanks to highly flexible RF components or to very high speed converters. Both have shown limitations and further research is needed to achieve highly flexible SDR. In fact, too major approaches emerge for designing a flexible RF: the first considers very large band RF that can therefore accommodate several systems, which suffers from bad sensitivity level though; the second relies on tuneable components with which parameters can be adapted to match the system requirements. These rules often contradict the guidelines RF designer are used to consider when defining an RF architecture, which is usually optimised for sensitivity, power consumption, and IC integration.

**VIII. GREEN COMMUNICATIONS**

Since the United Nations General Assembly in December 1987, and its Resolution 42/187, sustainable development has become an issue and an aspiration of our civilisation. The most often-quoted definition of sustainable development has been formulated by the Brundtland Commission [16] as the development that “meets the needs of the present without compromising the ability of future generations to meet their own need”. European Union in particular is continuously making efforts in legislation, regulations and recommendations to all sectors of the industry and agriculture, as well as in promoting ecology, and ecological life-style to address anticipated future climate issue and sustainable development in general. Surprising as it may seem, the ICT domain should be particularly concerned: currently, 3% of the world-wide energy is consumed by the ICT infrastructure, which causes about 2% of the world-wide CO₂ emissions. As the transmitted data volume increases approximately by a factor of 10 every 5 years, this corresponds to an increase of the associated energy consumption by approximately 20% per year, i.e., energy consumption is doubled every 5 years, which may cause serious problems in terms of sustainable development. A holistic approach to the wireless communication eco-sustainability is based on the idea that all wireless system and network components, as well as all infrastructure components, should contribute to the overall sustainable development of the considered ICT sector, and that the properties of that sector related to the impact on the eco-system cannot be determined only by its component parts alone. In other words, the eco-sustainability of the wireless network is not only about just the network elements, rather being about the whole network development, which would take all the relations and interactions between the network components to achieve the wireless communication sector sustainable development. Thus, a great attention has to be given to multiple network components and their dependencies to come out with noticeable power savings of the communication network.

Two major axes are to be distinguished within Green Communications:

- **Infrastructure**: this includes the transmission over the wired core network, switching, network management and...
maintenance, and is mostly concerned with the energy-savings matters.

- **Radio access**: energy consumption is still of importance for wireless devices, their users and their environment, and one should distinguish between on the one hand signal processing, RF front-end processing and amplification, and HW power supply (which has an impact on the global energy-consumption) and on the other energy radiation (which may have an impact on human health and the so-called electromagnetic pollution).

As mentioned above, the concept of Green Communications results from the emerging global consciousness concerning the ecosystem. It is noticeable that notions such as consciousness, intelligence, cognition, sensing, so far closely related to the human nature, are being used for the description of artificial intelligence and for the functionalities of various decision-making and learning devices, including designs for modern wireless communication. CR is such an emerging concept of an intelligent communication design that can orient itself in the radio environment, create plans, decide and take actions. One believes that consciousness is still related to a human being only, and it is an umbrella above the notion of cognition. It seems however that cognition alone of the CR devices can be used for making the Green Communications the reality at the radio access level. In fact, many researchers look at CR as an enabling technology for reasonable utilisation of radio resources, namely power consumption and available frequency band. Moreover, in this context, the “classical” understanding of green communications (i.e., power consumption) can be extended towards many other aspects of radio communications.

In order to achieve these expectations, research work has to be done into the following domains:

- **Signal processing techniques**: in order to sense the environment, and apply energy-optimised signal processing and transmission procedures.
- **Computer science**: encompassing decision making, learning, and prediction, in order to make the adequate decisions.
- **Hardware and software reconfigurability**: capabilities to modify and adapt the equipment behaviour and its power consumption.

The concept of a smart grid to deliver electricity from suppliers to consumers using digital technology is relatively new, and aims at controlling appliances at consumers’ homes and premises to save energy. Additionally, it is supposed to reduce cost and increase reliability and transparency of both the suppliers and the consumers. For the idea of smart power grids to happen, the need of robust, scalable and reliable communications infrastructure has been identified. It is supposed to be based on cooperative communications in home-area networks connected through wide-area networks. The privacy and safety of this communication is also a major issue.

One of the challenges of the future wireless radio systems is to globally reduce the electromagnetic radiation levels in order to reduce the wireless networks capacity are contributing to Green Communications. A few examples of the trendy topics include: MIMO processing, near-Shannon limit channel-coding schemes, and cooperative communications and relying. Thus, all the signal processing research lead in this sense in the next 20 years will be contributing to the Green Communication field.

Moreover, new adaptive spectrum sharing models should be designed, and cooperative communication techniques at all layers can be applied to increase the spectral and power efficiency, to enhance network coverage, and to reduce the outage probability in wireless networks. A result, one may expect better spectrum usage in available bands, release of some other bands that should be reserved for other purposes, e.g., radiation detection of various natures, lower transmission power, thus, lower interference. Cognitive and cooperative communications with all their aspects of the radio environmental awareness and flexible and efficient usage of resources are long term objectives.

In the sense of sustainable development, not only the energy consumption is considered, but also the spectrum pollution with the increase of radiation related to communication systems should be addressed. Spectrum may be considered as a natural and public resource, which should be carefully used, shared world-wide, and economised. Spectrum indeed only exists if there is sufficient energy to generate the electromagnetic waves, thus, spectrum savings is a question of capacity versus power spectral density of radio transmission.

Another challenge of future wireless radio systems is to globally reduce the electromagnetic radiation levels in order to reduce human exposure to radiation.

The power-efficiency of the radio hardware equipment cannot be overestimated as its contribution to the mobile-terminal, base-station and overall-network power consumption is noticeable. The silicon components optimisation towards lower power consumption in Watts per Gigabit of the data processing is recognised as the continuously hot research topic. Moreover, handset system design is a continuing area of innovation (examples include hardware design to improve music playback time). Nevertheless, it has been recognised, that not enough attention is paid to the behaviour of software taking into account an increasing number of applications relying on web services.

**IX. Conclusions**

This paper identifies medium-long term research challenges in wireless communications, based on the 1st NEWCOM++ Vision Book [1]. Section II addresses wireless networks, with emphasis on heterogeneity and opportunism. The basic ingredients allowing to successfully coping with the increasing demand of traffic and bit are addressed, via the increase of the spectral efficiency, increase of the number of base stations to provide a service in a given area; and increase of the total available bandwidth. Trends towards decentralisation and flat architectures, and perspectives on flexible spectrum management are shown as well. Self-organising networks are included in the analysis, since it is expected that many mechanisms will be established to allow the self-configuration of a
network with minimum human intervention. The features to be included in such an approach include: self-configuration, self-planning, self-optimisation, self-managing, and self-healing. Opportunistic networks are also addressed, linking with Delay Tolerant Networks (among others), vehicular communication networks are analysed, linking with Intelligent Transportation Systems, and the Future Internet is viewed from the Internet of Things and RFID perspectives.

Section III also addresses wireless networks, but focusing on cognition and cooperation, and providing a view on efficient ways of setting networks. When addressing Cognitive Radio Networks, several aspects of spectrum usage and management are discussed. A number of techniques to be developed for the implementation of efficient spectrum usage through cognitive radio networks are dealt with: spectrum sensing, spectrum management, spectrum mobility, and spectrum sharing mechanisms. Cognitive Positioning is then addressed, in relation with cognitive radio, Multi Carrier systems being taken as an example. Finally, the concept of Docitive Radios & Networks is introduced, i.e., a novel framework on radios & networks that teach other radios & networks. These radios & networks are not (only) supposed to teach them the end-result, but rather elements of the methods to getting there.

Section IV presents the main facets of research challenges that lie ahead towards the goal of understanding the ultimate performance limits of networks, and of designing innovative techniques to approximate and even achieve them. Various optimisation- and control theory driven techniques where put forward such as distributed optimisation and the max-weight control principle. Novel and disruptive approaches are described, such as network coding, and their potential is analysed. Furthermore, mathematical tools such as cooperative and non-cooperative game theory, learning theory and stochastic geometry will be needed in order to model and understand the spontaneous interactions of massively dense autonomic networks. Clearly, the way ahead is promising. Novel and disruptive approaches will need to be undertaken, oftentimes relying on cross-disciplinary techniques migrating from a wide range of disciplines and thrusts as described above.

Section V discusses theoretical and practical problems on bandwidth- and energy-efficient radio access. The first part of the section discusses on approaching the fundamental communication limits, by addressing them, and then, by showing some techniques that enable this approach, namely coding/decoding techniques, as well as adaptive modulation and coding. Afterwards, iterative techniques in wireless receivers are presented, including the framework for iterative processing, applications on iterative receiver design, and low complexity decoding. Finally, cognitive spectrum usage (always for the purpose of advancing spectral utilisation) with a special focus on applied game theory is discussed, including multicarrier techniques and distributed spectrum usage.

Section VI addresses on multimedia signal processing for wireless delivery, encompassing a discussion on the current situation, a review of enabling technologies, and a view into the global optimisation problem. A discussion on the general trend is presented, including some of the current technologies, presenting some of the driving forces, and extracting general consequences for future work. The review on the enabling technologies addresses 3-D representation (based on stereo vision and signal processing), screens and cameras, the network (e.g., mobile ad hoc networks and network coding, and possible research direction). The global optimisation issue includes cross-layer optimisation, joint source-channel-network-protocol coding, and services and usage from a wireless multimedia processing viewpoint.

Section VII fits within recent trends in silicon technology and communication system demands, which exhibit a growing gap between application needs and what the technology can deliver. A key driver for the telecom industry is the mobile business. Mobility, which relies on battery operated handheld devices, provide stringent requirements on equipments in terms of processing power, power consumption and flexibility. At the same time the battery and silicon technology does not progress at the same pace. The emergence of new standards implementing ever more efficient air interfaces also put stringent constraints on the design time. Thus, the reuse of hardware building blocks and a proper methodology and tools are needed to evaluate hardware performance tradeoffs at the earliest stage. Flexible radio is a promising approach in this regard. However, a unified framework is still to be found to enable the co-design of communication functions and hardware platforms.

Section VIII deals with “green” communications, a concept drawing from the universally shared concern about climate changes and aspiration toward a sustainable development guided by ecological considerations. European Union is taking efforts to address the climate issues including the ICT related issues and footprints on the environment. It is of major importance for the sake of radio domain in the 21st century to bring progress to people and to their confidence, and not to make them fear of the radio evolution. Green Communications developments should provide this confidence as an ICT approach for finding solutions to the issues of CO₂ emission, energy consumption, resources utilisation, electromagnetic pollution and heath issues. Undoubtedly, this can be considered as a challenge for the next 20 years.

Acknowledgement

This paper is based on the 1st NEWCOM++ Vision Book, to which many project participants have contributed. The authors would like to acknowledge all colleagues that provided inputs to this report, and hence, indirectly, to this paper.

References


Sergio Benedetto is a professor at Politecnico di Torino. Active for more than 30 years in the field of digital communications, he has coauthored 4 books and over 250 papers. He has received the “Premio Siemens per le Telecomunicazioni” in 1993 and 1995, the “Italgas International Prize for Research and Technological Innovation” in 1998, the “Cristoforo Colombo International Award for Communications” in 2006, and the “IEEE Communications Society Edwin Howard Armstrong Award” in 2008. He has been Chair of the Communication Theory Technical Committee, was TP Chair of the Communication Theory Symposium at ICC 2000 and ICC 2006, and General Chair of the Communication Theory Workshop in 2004. An IEEE Fellow, Sergio has been Area Editor for IEEE Transactions on Communications and a Distinguished Lecturer of ComSoc. He has been Vice President of Technical Activities of ComSoc in 2006-2007, the Vice-President for Publication in 2008-2009, and is now Vice President for Member Relations. He is member of the Turin Academy of Sciences.

Luis M. Correia was born in Portimao, Portugal, on October 1958. He received the Ph.D. in Electrical and Computer Engineering from IST-TUL (Technical University of Lisbon) in 1991, where he is currently a Professor in Telecommunications, with his work focused in Wireless/Mobile Communications in the areas of propagation, channel characterisation, radio networks, traffic, and services. He has acted as a consultant for Portuguese mobile communications operators and the telecommunications regulator, besides other public and private entities. Besides being responsible for research projects at the national level, he has been active in various ones within European frameworks of RACE, ACTS, IST, ICT and COST (where he also served as evaluator and auditor), having coordinated two COST projects. He has been supervising students at both the M.Sc. and Ph.D. levels, having authored more than 270 papers and communications in international and national journals and conferences, for which he has served also as a reviewer, editor, and board member. He was part of the COST Domain Committee on ICT. He was the Chairman of the Technical Programme Committee of several conferences, namely PIMRC’2002. He is part of the Expert Advisory Group and of the Steering Board of the European eMobility platform, and the Chairman of its Working Group on Applications.
Spectrum Occupancy in Realistic Scenarios and Duty Cycle Model for Cognitive Radio

Miguel López-Benítez and Fernando Casadevall

Abstract—Most of the spectrum occupancy measurement campaigns performed to the date in the context of cognitive radio are based on measurements in outdoor high points such as building roofs, balconies and towers. Although these measurement scenarios enable a more accurate estimation of the primary transmitters’ spectral activity, they may not be representative of the spectrum occupancy perceived by a cognitive radio user in many other interesting practical situations where users are not placed in a static high point. In this context, this work presents the results obtained in a spectrum measurement campaign performed over a rich diversity of measurement scenarios of practical interest. The considered scenarios include not only high points but also indoor environments as well as outdoor locations at the ground level in open areas and between buildings. The variety of considered measurement scenarios provides a broader view and understanding of dynamic spectrum occupancy under different practical scenarios of interest. The impact of considering various locations on the spectral activity perceived by a cognitive radio user is determined, analyzed and quantified. Moreover, a theoretical model for the occupancy levels observed at different locations is developed and verified with the obtained results.

Index Terms—Cognitive radio; Dynamic spectrum access; Measurement campaign; Spectrum occupancy; Duty cycle model.

I. INTRODUCTION

The owned spectrum allocation policy in use dates from the earliest days of modern radio communications. This scheme has been proven to effectively control interference among different wireless systems and simplify the design of hardware for use at a known and fixed radio frequency range. However, with the rapid proliferation of new operators, innovative services and wireless technologies during the last decades, the vast majority of the available spectrum regarded as usable has already been occupied. Some recent spectrum measurements have demonstrated however that most of spectrum, though allocated, is vastly underutilized. In this context, the Cognitive Radio (CR) paradigm has emerged as a promising solution to conciliate the conflicts between spectrum demand growth and underutilization without changes to the existing legacy wireless systems.

The basic underlying idea of CR is to allow unlicensed users to access in an opportunistic and non-interfering manner some licensed bands temporarily unoccupied by licensed users. The operating principle is to identify spatial and temporal spectrum gaps not occupied by licensed (primary) users, usually referred to as spectrum holes or white spaces, place unlicensed (secondary) transmissions within such spaces and vacate the channel as soon as primary users return. Secondary unlicensed transmissions are allowed following this operating principle as long as they do not cause excessive, harmful interference levels to the primary network.

The amount of transmission opportunities available to the secondary system depends on the degree to which allocated spectrum bands are used by the primary system. In the real world, the occupancy level of any spectrum band can be quantitatively determined by means of field measurements. Measurements of the radio environment can provide valuable insights into current spectrum usage. A proper understanding of current spectrum usage can be extremely useful, not only to policy makers in order to define adequate Dynamic Spectrum Access (DSA) policies for improving the exploitation of the currently underutilized spectral resources, but also to the research community in order to develop spectrum usage models and identify the most suitable and interesting frequency bands for the future deployment of the CR technology.

To the date, several spectrum measurement campaigns covering wide frequency ranges [1], [2], [3], [4], [5], [6] as well as some specific licensed bands [7], [8], [9], [10], [11] have already been performed in diverse locations and scenarios in order to determine the degree to which allocated spectrum bands are used in real wireless communication systems. However, most of previous spectrum occupancy studies are based on measurements performed in outdoor environments and more particularly in outdoor high points such as building roofs, balconies and towers. The main advantage of high points is that they provide direct line-of-sight to many kinds of important transmitters and therefore enable a more accurate measurement of the spectral activity. Nevertheless, this scenario may not be representative of the spectrum occupancy perceived by a secondary network in many other interesting practical situations where the secondary antenna is not placed in a static high point (e.g., a mobile user communicating inside a building or while walking in the street between buildings). The measurement of real network activities in additional scenarios of practical significance is therefore required for an adequate and full understanding of the dynamic use of spectrum.

In this context, this work presents the results obtained in a spectrum measurement campaign performed over a rich diversity of measurement scenarios of interest in a densely populated urban environment in the city of Barcelona, Spain. The considered scenarios include not only high points but also indoor environments as well as outdoor locations at the ground level in open areas and between buildings. The variety of
considered measurement scenarios provides a broader view and understanding of dynamic spectrum occupancy under different practical scenarios of interest. The aim of this work is to analyze, determine and quantify the impact of considering various locations on the spectral activity perceived by a secondary user with respect to that observed in an outdoor high point. This information will be very useful in the analysis of realistic scenarios as well as the development of simulation tools and theoretical studies. Furthermore, a theoretical model for the occupancy levels observed at different locations is developed and verified with the obtained results. This model can be used to predict the occupancy level of a given spectrum band perceived at any geographical location based on the knowledge of some simple signal parameters.

II. MEASUREMENT SETUP AND METHODOLOGY

The measurement configuration employed in this work (see Fig. 1) relies on a spectrum analyzer setup where different external devices have been added in order to improve the detection capabilities and hence obtain more accurate and reliable results. The design is composed of two broadband discone-type antennas covering the frequency range from 75 to 7075 MHz, a Single-Pole Double-Throw (SPDT) switch to select the desired antenna, several filters to remove undesired overloading and out-of-band signals, a low-noise preamplifier to enhance the overall sensitivity and thus the ability to detect weak signals, and a high performance spectrum analyzer to record the spectral activity. The measurement setup and methodology employed in this work have been carefully designed based on the findings of the study presented in [12], where some important methodological aspects to be accounted for when evaluating spectrum occupancy in the context of CR are analyzed and discussed. A detailed description of the measurement setup configuration parameters and measurements setup design principles as well as the methodological procedures considered in this study are in [12], [13], [14].

III. MEASUREMENT SCENARIOS

The scenarios defined for this measurement campaign include not only outdoor but also indoor locations. Measurements in indoor locations provide information about the spectral activity that would be perceived by secondary users operating in indoor environments. Similarly, outdoor measurements give us some insights into the spectral activity that would be perceived by secondary users operating in outdoor environments, at various physical locations of practical interest. For the measurements in outdoor locations, three different kinds of scenarios have been considered, namely high points, narrow streets and open areas. Measurements in high points provide reliable information about the actual spectral occupancy patterns of several primary transmitters, while narrow streets and open areas give us an idea of the perception of secondary users moving within a urban environment with different levels of radio propagation blocking.

For indoor experiments, the measurement equipment was placed inside an urban building, in the middle floor of a three-floor building belonging to the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. For outdoor high point measurements, the equipment was placed in the roof of the same building (latitude: 41º 23’ 20” north; longitude: 2º 6’ 43” east; altitude: 175 meters). The selected place is a strategic location with direct line-of-sight to several transmitting stations located a few tens or hundreds of meters away from the antenna and without buildings blocking the radio propagation. This strategic location enabled us to accurately measure the spectral activity of, among others, TV and FM broadcast stations, several nearby base stations for cellular mobile communications and a military head quarter as well as some maritime and aeronautical transmitters due to the relative proximity to the harbor and the airport. For measurements in narrow streets and open areas, the measurement equipment was moved within the UPC’s Campus Nord. The different geographical locations considered in this measurement campaign are illustrated in Fig. 2, where an aerial view of the campus is shown. The description of the measurement locations is provided in Table I.

![Fig. 1. Measurement setup employed in this study.](image1)

![Fig. 2. Measurement locations in urban environment.](image2)

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>DESCRIPTION OF THE MEASUREMENT LOCATIONS.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measurement point</strong></td>
<td><strong>Environment</strong></td>
</tr>
<tr>
<td>1</td>
<td>Outdoor high point (building roof)</td>
</tr>
<tr>
<td>2</td>
<td>Indoor (building room)</td>
</tr>
<tr>
<td>3 – 7</td>
<td>Outdoor at ground level in narrow streets</td>
</tr>
<tr>
<td>8 – 10</td>
<td>Outdoor at ground level between buildings</td>
</tr>
<tr>
<td>11 – 12</td>
<td>Outdoor at ground level in open areas</td>
</tr>
</tbody>
</table>
IV. SPECTRUM OCCUPANCY RESULTS

A. Occupancy Metrics

The occupancy level of various spectrum bands is quantified throughout this work in terms of the duty cycle. This section provides a formal definition for such occupancy metric.

The duty cycle is computed based on a finite set of discrete measurements collected along a range of frequencies $F_{\text{span}} = F_{\text{stop}} - F_{\text{start}}$ (frequency span) and over a period of time $T_{\text{span}} = T_{\text{stop}} - T_{\text{start}}$ (time span).

The measured discrete time instants $t_i$ ($T_{\text{start}} \leq t_i < T_{\text{stop}}$) are given by
\[
t_i = T_{\text{start}} + (i - 1) \cdot T_r, \quad i = 1, 2, \ldots, N_t
\]
where $T_r$ represents the time resolution and is determined by the spectrum analyzer’s sweep time, which in turn depends on the selected configuration parameters. For a certain time resolution $T_r$, the number of traces $N_t$ collected within a time span $T_{\text{span}}$ is given by $N_t = T_{\text{span}}/T_r$.

The measured discrete frequency points $f_j$ ($F_{\text{start}} \leq f_j < F_{\text{stop}}$) are given by
\[
f_j = F_{\text{start}} + (j - 1) \cdot F_r, \quad j = 1, 2, \ldots, N_f
\]
where the frequency resolution $F_r = F_{\text{span}}/N_f$ is the frequency bin, determined by the selected frequency span $F_{\text{span}}$ and the number of frequency points $N_f$ measured by the spectrum analyzer.

The set of Power Spectral Density (PSD) samples collected by a spectrum analyzer over a time span $T_{\text{span}}$ and along a frequency span $F_{\text{span}}$ can be represented by a $N_t \times N_f$ matrix $M$ as
\[
M = [M(t_i, f_j)]
\]
where each matrix element $M(t_i, f_j)$ represents the PSD sample captured at time instant $t_i$ ($i = 1, 2, \ldots, N_t$) and frequency point $f_j$ ($j = 1, 2, \ldots, N_f$).

To compute the duty cycle, the presence or absence of a licensed signal needs to be determined for each PSD sample $M(t_i, f_j)$. In other words, for each captured PSD sample it is necessary to determine whether the sample corresponds to a licensed signal sample or a noise sample. Several signal detection principles have been proposed in the literature to perform such task [15]. They provide different trade-offs between required sensing time, complexity and detection capabilities. Depending on how much information is available about the signal used by the licensed network different performances can be reached. However, in the most generic case no prior information is available. If only power measurements of the spectrum utilization are available, the energy detection method is the only possibility left. Due to its simplicity and relevance to the processing of power measurements, energy detection has been a preferred approach for many past spectrum studies. Energy detection compares the received signal energy in a certain frequency band to a predefined decision threshold. If the signal energy lies above the threshold, a licensed signal is declared to be present. Otherwise, the measured frequency channel is supposed to be idle. Following this principle, a binary spectral occupancy matrix
\[
\Omega = [\Omega(t_i, f_j)]
\]
is defined, where each element $\Omega(t_i, f_j) \in \{0, 1\}$ is computed as
\[
\Omega(t_i, f_j) = u(M(t_i, f_j) - \lambda_j)
\]
with $\lambda_j$ being an energy decision threshold for frequency point $f_j$ and $u(\cdot)$ is the unit (Heaviside) step function, thus resulting in
\[
\Omega(t_i, f_j) = \begin{cases} 
0, & M(t_i, f_j) - \lambda_j < 0 \Rightarrow M(t_i, f_j) < \lambda_j \\
1, & M(t_i, f_j) - \lambda_j \geq 0 \Rightarrow M(t_i, f_j) \geq \lambda_j 
\end{cases}
\]

Each element $\Omega(t_i, f_j)$ in matrix $\Omega$ represents the presence $\Omega(t_i, f_j) = 1$ or absence $\Omega(t_i, f_j) = 0$ of a licensed signal at time instant $t_i$ and frequency point $f_j$, according to the energy detection principle based on an energy decision threshold $\lambda_j$. The decision threshold is set according to the Probability of False Alarm 1% (PFA 1%) criterion as explained in [12].

For each measured frequency point $f_j$, the duty cycle $\Psi_j$ is computed as the fraction of PSD samples, out of all the PSD samples recorded at that frequency, that lie above the decision threshold $\lambda_j$ and hence that are considered as samples of occupied channels:
\[
\Psi_j = \frac{1}{N_t} \sum_{i=1}^{N_t} \Omega(t_i, f_j)
\]
For a frequency point $f_j$, this metric represents the fraction of time that the frequency is considered to be occupied. For a certain frequency span (i.e., range of frequencies $j = 1, 2, \ldots, N_f$), the average duty cycle $\Psi$ of the band is computed by averaging the duty cycle $\Psi_j$ of all the $N_f$ frequency points measured within the band:
\[
\Psi = \frac{1}{N_f} \sum_{j=1}^{N_f} \Psi_j = \frac{1}{N_t N_f} \sum_{i=1}^{N_t} \sum_{j=1}^{N_f} \Omega(t_i, f_j)
\]
This metric represents the average degree of spectrum utilization within certain time ($T_{\text{span}}$) and frequency ($F_{\text{span}}$) spans. The duty cycle is usually given in percentage and this is the convention adopted in this study.

B. Comparison of Location 1 and Location 2

Most of previous spectrum occupancy studies are based on measurements performed in outdoor high points such as building roofs, balconies and towers\(^1\). This section presents and analyzes the results obtained in a urban indoor environment (location 2) taking as a reference the results obtained in an outdoor high point (location 1). The aim of this section is to determine the impact of considering an indoor environment on the spectral activity perceived by a secondary user with respect to that observed in an outdoor high point. Although the measurement conditions considered in both cases were identical, the time instants were different (i.e., both locations were indoors). This circumstance introduces some random component in the obtained results since different transmissions were present in each case. However, it is worth noting that the aim of this section is not to characterize the instantaneous spectrum occupancy in the time domain but the

\(^1\)In our measurement campaign, this scenario corresponds to location 1.
average occupancy rate from a statistical point of view. For sufficiently long measurement periods as those considered in this study (24 hours), the impact of different instantaneous transmissions is averaged and the obtained average duty cycle value can be considered as a representative indication of the spectral activity in the bands under study.

From a qualitative point of view, the results obtained in location 2 follow the same trend as in location 1\textsuperscript{2}, with higher occupancy rates at lower frequencies. As it can be appreciated in Table II, the average spectrum occupancy is moderate below 1 GHz and very low above 1 GHz. The significantly lower average occupancy rates observed in Table II for the indoor location can be explained by the fact that most of wireless transmitters are located outdoor and the propagation loss due to outdoor-indoor signal penetration leads to lower signal strengths in the indoor scenario, which in turn results in lower occupancy rates. In principle, the lower average duty cycles obtained for the indoor case suggest the existence of a higher amount of free spectrum. However, this result should be interpreted carefully, taking into account the specific circumstances of particular bands.

To analyze the impact of the indoor location on the occupancy rate for various specific bands, it is convenient to distinguish four different possible cases according to the location of transmitters and receivers, as shown in Table III. Based on this classification, the results for various bands of interest are shown in Fig. 3. Notice that for certain bands the classification is not straightforward. For example, in the downlink direction of cellular mobile communication systems the receivers are mobile users that may be located indoor and outdoor simultaneously. In practice, it is not possible to reliably determine the location of every transmitter and receiver operating in a certain band, which results in some uncertainty. In spite of that, some general trends can be inferred from the results shown in Fig. 3.

For bands allocated to systems where the transmitters are always outdoor (cases I/II), the indoor duty cycles are in general notably lower, as expected, due to the outdoor-indoor signal penetration loss. In case I (systems with outdoor receivers) the lower indoor occupancy rates indicate the availability of more free spectrum, since an indoor secondary user transmitting in channels sensed as free would not cause harmful interference to primary indoor receivers. However, in case II (systems with indoor receivers) the lower indoor duty cycles do not necessarily imply the existence of more white spaces, since in this case transmitting in a channel sensed as unoccupied could potentially result in interference to primary indoor receivers.

For bands allocated to systems with indoor transmitters (cases III/IV), in general the average duty cycles tend to be higher in the indoor location (with some exceptions as the E-GSM 900 uplink band, which might be due to the presence of outdoor transmitters in such band). Following a similar argument, in case IV (systems with indoor receivers) this indicates the availability of a lower amount of free spectrum, while it could not be necessarily the situation in case III (systems with outdoor receivers). In any case, the differences observed in this experiment between outdoor and indoor occupancy rates in cases III/IV are not as significant as in cases I/II.

In summary, although average duty cycles tend to be lower in indoor locations (with some unimportant particular exceptions), this does not necessarily indicate the existence of more free spectrum. The particular circumstances of the specific bands being sensed and the characteristics of the systems operating over them need to be carefully considered before declaring a band as truly available for a potential secondary usage; otherwise, harmful interference could be caused to the primary licensed system.

\textsuperscript{2}A detailed analysis of the occupancy results obtained for location 1 can be found in [14].
Based on the previous results, from a practical point of view it can be stated that the output of spectrum sensing procedures is not enough to declare a band as truly available for secondary access. Some additional techniques may be required such as, for example, sensing both the uplink and downlink directions of FDD-based systems in order to guarantee that the channel can be accessed opportunistically, or employing signal processing techniques as the one described in [14] in order to determine the signal standard present in a certain band before deciding whether a band may be accessed without inducing harmful interference.

C. Comparison of Location 1 and Locations 3-12

This section presents and analyzes the results obtained in urban narrow streets and open areas (locations 3-12) taking as a reference the results obtained in an outdoor high point (location 1). The aim of this section is to determine the impact of considering different outdoor locations at the ground level on the spectral activity perceived by a secondary user with respect to that observed in an outdoor high point. The locations under study in this section can be considered as a representative scenario for secondary mobile users communicating while walking on the street in an urban environment.

As in section IV-B, each location has been measured at a different time instant. As it has been mentioned above, the random component introduced by the presence of different transmissions at different times could be averaged by considering a sufficiently long measurement period. However, since the presence of an operative was required in the measurements, periods of 24 hours as in location 1 were infeasible and were therefore shortened to 1 hour in locations 3-12. To reduce the impact of random components and make the results of locations 1 and 3-12 comparable, the average duty cycle obtained in locations 3-12 has been normalized by the average duty cycle in location 1 obtained when considering the samples corresponding to the same time interval. Therefore, if an average duty cycle $\Psi_k$ is obtained for location $k$ ($k = 3, 4, \ldots, 12$) based on the samples captured during a 1-hour interval between time instants $T_{\text{start}}$ and $T_{\text{stop}}$, the samples captured at location 1 between the same $T_{\text{start}}$ and $T_{\text{stop}}$ values are used to compute an average duty cycle $\Psi_1$. The normalized average duty cycle for location $k$ is then obtained as $\Psi_k^* = \Psi_k / \Psi_1$. This procedure reduces the randomness of the obtained results and enables a fairer comparison between the outdoor high point and the rest of outdoor positions.

For most of the bands and locations measured in this experiment the obtained normalized average duty cycle is lower than one, meaning that the average duty cycle measured in different locations at the ground level is in general lower than in high points. This is a consequence of the radio propagation blocking caused by buildings and other obstacles: under non-line-of-sight conditions, the direct ray (i.e., the strongest signal component) is lost; only multi path propagation components attenuated by reflection, refraction and diffraction are received, thus resulting in lower received signal levels and therefore in lower average duty cycles. From a practical point of view, this indicates that a secondary user at the ground level would perceive an amount of white space higher than that predicted by measurements performed in high points. Nevertheless, it is worth highlighting that this should be interpreted carefully, taking into account the specific circumstances of each band. In the following, some particular bands of interest are discussed. The obtained results are shown in Fig. 4.

Fig. 4 shows the spatial distribution of the normalized average duty cycle for the TV, UMTS downlink, E-GSM 900 downlink and DCS 1800 downlink bands, respectively. The common feature of these bands is that the transmitters are located outside the region under study. In the TV band, it can be clearly appreciated that the normalized average duty cycle is lower in closed regions. Thus, in locations 4 and 6, where radio propagation blocking caused by buildings is more intense, its value is lower than in other less closed regions such as locations 3, 5 and 7. A similar trend is observed for UMTS, E-GSM 900 and DCS 1800 downlink bands. In the two last cases, however, location 5 constitutes an exception, which could be explained by the different relative
position of transmitters and by the fact that the same physical scenario may result in very dissimilar propagation scenarios at different frequencies. Comparing locations 8, 9 and 10, the deepest region (location 9) exhibits the lowest normalized average duty cycle in the case of TV, as expected, but the highest value in the case of the cellular mobile communication bands, which could be explained by the use of micro-cells and repeaters in shadowed regions as location 9. Regarding locations 11 and 12, it is interesting to note that a higher spectral activity level was recorded in location 11 despite the presence of some surrounding buildings with respect to the open region in location 12. The detection by the measurement equipment of some additional signal components reflected in such buildings could explain the recording of higher activity levels in a less open region. Although lower duty cycles have been observed at the ground level in the TV, UMTS, E-GSM and DCS 1800 downlink bands, it is worth noting that this does not necessarily imply the existence of more opportunities for secondary access. As a matter of fact, some faded primary signals might be undetected at the ground level due to blocking buildings and other obstacles, in which case an exceptionally harmful interference would be caused to the intended primary receivers, who in these bands would be operating in the proximity of the secondary users. These experimental results highlight the importance of detection sensitivity in secondary networks and suggest the need of some additional techniques as mentioned in section IV-B.

In the previous bands, where the transmitters were outside the region under study, the results have shown some general occupancy trends. However, for other bands with transmitters operating inside the region under study, no particular occupancy trends have been observed. In this case, the obtained results might depend not on the actual spectral usage of such bands but rather on the random and fluctuating geographical distribution of transmitters inside the considered region.

The main conclusion derived from the obtained results is that the spectral activity perceived by a secondary user for a certain band in realistic urban scenarios strongly depends on the user location, with significant variations even in physical areas as reduced as the one considered in this study (≈ 180 × 260 m). This indicates that the conclusions derived from high point measurements may not be well suited to other realistic outdoor scenarios with users at the ground level.

V. DUTY CYCLE MODEL FOR COGNITIVE RADIO

In this section, a model describing the spectral occupancy level perceived at various geographical locations is developed and verified. This model can be used to predict the spectral occupancy, expressed in terms of the duty cycle, that would be perceived by a CR user at any location based on the knowledge of some simple signal parameters. In this section we focus on modeling the duty cycle for TV bands. The interest of considering this band is twofold. On one hand, the first real CR deployments are expected in the TV bands following the IEEE 802.22 standard [16]. Therefore, the development of spectrum usage and prediction models for TV bands is more necessary than for other bands. On the other hand, TV stations are constant-power transmitters with a 100% activity factor, which provides the basis for simple occupancy models that can be extended in the future for non-constant-power transmitters and/or discontinuous transmission patterns.

The model relies on the premise that the average noise channel power and the received average signal channel power can be adequately modeled as Gaussian random variables, which is demonstrated in Fig. 5 and 6, respectively. For a given primary Radio Frequency (RF) channel, the average channel power at time instant $t_i$ is computed as the integral (sum of discrete values) of the PSD levels measured at the frequency points $f_j$ comprised within the RF channel limits, and within a certain time period around $t_i$. Average channel powers are therefore obtained by averaging the power levels measured at different time instants and frequency points. Hence, the Central Limit Theorem can be employed to explain the Gaussian behavior of Fig. 5 and 6. As shown in Fig. 5, the average noise channel power measured at different TV channels increases with the channel number (i.e., the frequency), but can always
be modeled as a Gaussian random variable. Fig. 6 shows that the average signal channel power can also be modeled as a Gaussian random variable. Although the Gaussian curve corresponding to the signal’s mean and standard deviation not always provides a perfect fit, it is able to reasonably describe the distribution of the received signal levels.

As stated in section IV-A, the duty cycle represents the fraction of time that a channel is occupied. Following the energy detection principle, a channel is considered to be occupied whenever the average signal channel power is greater than a decision threshold \( \lambda \). Assuming that the average signal channel power follows a Gaussian law with mean \( \mu_S \) and standard deviation \( \sigma_S \), the duty cycle \( \Psi \) can be computed (see Fig. 7) as:

\[
\Psi = \frac{1}{\sqrt{2\pi} \sigma_S} \int_{\lambda}^{\infty} e^{-\frac{(x-\mu_S)^2}{2\sigma_S^2}} \, dx \tag{9}
\]

Making the variable change \( z = \frac{1}{\sqrt{2}} \left( \frac{x-\mu_S}{\sigma_S} \right) \), the integral is readily obtained to be:

\[
\Psi = \frac{1}{\sqrt{2}} \int_{\lambda}^{\infty} e^{-\frac{z^2}{2}} \, dz = \frac{1}{\sqrt{2}} \text{erfc} \left( \frac{\lambda - \mu_S}{\sqrt{2} \sigma_S} \right) \tag{10}
\]

\[
\Psi = Q \left( \frac{\lambda - \mu_S}{\sigma_S} \right) \tag{11}
\]

where \( \text{erfc}(\cdot) \) is the complementary error function and \( Q(\cdot) \) is the Gaussian \( Q \)-function.

The decision threshold \( \lambda \) in equation 12 can be established according to various methods. The study performed in [12], where some of them were comparatively evaluated and analyzed, concluded that the PFA 1% method represents an adequate trade-off between ability to detect the presence of weak signals and detection errors. In the context of signal detection, the PFA is defined as the probability that a signal is declared to be present in a certain channel when the channel is actually free. For an energy detector, this occurs when the noise power is greater than the decision threshold \( \lambda \). For a desired target PFA, denoted as \( P_{fa} \), the PFA method sets the decision threshold \( \lambda \) so that a fraction \( P_{fa} \) of noise power samples is allowed to lie above the decision threshold \( \lambda \). Assuming a Gaussian distribution for the average noise channel power with mean \( \mu_N \) and standard deviation \( \sigma_N \), this implies (see Fig. 7) that:

\[
\frac{1}{\sqrt{2\pi} \sigma_N} \int_{\lambda}^{\infty} e^{-\frac{(x-\mu_N)^2}{2\sigma_N^2}} \, dx = Q \left( \frac{\lambda - \mu_N}{\sigma_N} \right) = P_{fa} \tag{13}
\]

Solving in equation 13 for \( \lambda \) yields the decision threshold:

\[
\lambda = Q^{-1}(P_{fa}) \sigma_N + \mu_N \tag{14}
\]

where \( Q^{-1}(\cdot) \) denotes the inverse of \( Q(\cdot) \).

Substituting equation 14 into equation 12 finally yields the duty cycle model:

\[
\Psi = Q \left( \frac{Q^{-1}(P_{fa}) \sigma_N - \text{SNR}}{\sigma_S} \right) \tag{15}
\]

where \( \text{SNR} = \mu_S - \mu_N \) represents the average Signal-to-Noise Ratio (SNR) expressed in decibels, while \( \sigma_S \) and \( \sigma_N \) are the standard deviation of the signal and noise powers also in decibels.

To validate equation 15, Fig. 8 and 9 depict, for analogical and digital TV channels respectively, the empirical duty cycle obtained for the 12 measured locations in Fig. 2. The duty cycle is shown as a function of the difference between the measured average signal and noise powers in dBm, i.e. the SNR in dB. The dependence of the perceived spectral activity with the geographical location is reflected in the different SNR values observed at each location. The theoretical curve of equation 15 corresponding to the measured signal and power standard deviations is also shown for target PFA of 1% and 10%. As it can be appreciated, the model perfectly agrees with the empirical values for both analogical and digital channels. More interestingly, the fit is reasonably accurate also for those channels for which the Gaussian fit in Fig. 6 was not perfect. These results demonstrate that equation 15 is able to accurately predict the spectral activity that would be perceived by a CR user at any position based on the knowledge of some basic signal parameters.

VI. CONCLUSION

Most of previous spectrum occupancy measurement campaigns in the context of cognitive radio have been performed in outdoor high points. The results obtained in such measurement locations, however, are not representative of the spectrum occupancy perceived by a cognitive radio user in many other locations of practical interest. In this context, this work has presented the results obtained in a spectrum measurement campaign performed over a rich diversity of measurement scenarios. The obtained results indicate that the occupancy level perceived by secondary users strongly depends on the considered location, with significant variations even in reduced physical areas. A theoretical model for the occupancy levels observed at different locations has been developed and verified with the obtained results.

ACKNOWLEDGMENTS

The authors wish to acknowledge the activity of the Network of Excellence in Wireless COMMunications NEWCOM++ of the European Commission (contract n. 216715)
Fig. 8. Validation of duty cycle model for analogical TV channels.

Fig. 9. Validation of duty cycle model for digital TV channels.
that motivated this work. The support from the Spanish Ministry of Science and Innovation (MICINN) under FPU grant AP2006-848 is hereby acknowledged.

REFERENCES


Miguel López-Benítez (www.lopezbentez.es) received a B.Sc. degree (2003) and a M.Sc. degree (2006) in Telecommunications Engineering, both with First-Class Honors, from Miguel Hernández University (UMH), Elche, Spain, and was the recipient of the 2003 and 2006 University Education National Awards, a distinction from the Spanish Ministry of Education and Science to the best national academic records. His degree thesis was awarded a national research prize from the France Telecom foundation in the context of the 5th Archimedes University Competition, the main young scientist contest in Spain organized by the Spanish Ministry of Education and Science. From February 2003 to July 2006 he actively participated in several Spanish national and regional research projects in what today is the Ubiquitous Wireless Communications Research Laboratory (Uwicore, www.uwicore.umh.es), UMH. Since September 2006 he is with the Mobile Communication Research Group (GRCM, www.grcm.tsc.upc.edu) at the Department of Signal Theory and Communications of the Technical University of Catalonia (UPC), Barcelona, Spain. He currently is pursuing a Ph.D. degree in Telecommunications Engineering at UPC with the support of a FPU grant from the Spanish Ministry of Education and Science, and collaborates as a teaching assistant. He has been or is actively involved in European-funded projects AROMA, NEWCOM++, and FARAMIR along with Spanish project COGNOS. His current research interests include the field of mobile radio communication systems, with a special emphasis on radio resource management, heterogeneous wireless systems, quality of service provisioning, spectrum modeling and opportunistic/dynamic spectrum access in cognitive radio networks.

Fernando J. Casadevall (ferranc@tsc.upc.edu) received an Engineer of Telecommunication and Dr. Engineering degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1977 and 1983 respectively. In 1978 he joined UPC, where he was an Associate Professor from 1983 to 1991. He is currently Full Professor in the Signal Theory and Communications Department. After graduation he was concerned with equalization techniques for digital fibre optic systems. He has also been working in the field of digital communications with particular emphasis on digital radio and its performance under multipath propagation conditions. In the last fifteen years, he has mainly been concerned with the performance analysis and development of digital mobile radio systems. In particular, his research interests include cellular and personal communication systems, multipath transceiver design (including software radio techniques), mobility and radio resources management, and end-to-end QoS issues. During the last ten years he participated in more than thirty research projects funded by both public and private organizations. In particular, he actively participated in ten research projects founded by the European Commission, being the Project Manager of three of them: ARROWS, EVEREST and AROMA. (See http://www.grcm.tsc.upc.edu for details.) Prof. Casadevall has published around one hundred and fifty technical papers in both international conferences and magazines; most of them correspond to IEEE publications. He has also been a Technical Program Committee member of different international IEEE supported conferences, as well as a reviewer of several IEEE magazines. From October 1992 to January 1996 he was responsible for the Information Technology Area in the National Agency for Evaluation and Forecasting (Spanish Nation Research Council).
A Review of OFDMA and Single-Carrier FDMA and Some Recent Results

Cristina Ciochina and Hikmet Sari, Fellow, IEEE

Abstract—The controversial debate on OFDM vs. single-carrier (SC) transmission started back in the 1980s at the time of the European Digital Audio Broadcasting (DAB) and Digital Video Broadcasting (DVB) projects. The same debate took place in wireless communications a decade later, and OFDM transmission with TDMA was adopted in the IEEE 802.11a specifications for wireless local area networks (WiFi) and by the WiMAX Forum for fixed WiMAX systems. Later, orthogonal frequency-division multiple access (OFDMA) was adopted by the WiMAX Forum for mobile WiMAX systems and more recently by the 3GPP for the downlink of Long Term Evolution (LTE) systems. In contrast, single-carrier FDMA was adopted for the uplink of LTE. In this overview paper, we will review these historic developments and give some recent results on OFDMA and Single-Carrier FDMA.

Index Terms—OFDM, OFDMA, SC-FDMA.

I. INTRODUCTION

Orthogonal frequency-division multiplexing (OFDM) which was known since the 1950s, was revived in the 1980s with the European Digital Audio Broadcasting (DAB) [1] and Digital Video Broadcasting (DVB) projects. This technique was standardized for both DAB and digital terrestrial TV broadcasting (DVB-T). The technical literature at that time, mostly by authors involved in the DAB and the DVB projects, did not leave much alternative to using OFDM for digital terrestrial TV, particularly for mobile reception.

In 1993, Sari et al. presented a conference paper [2], which reviewed the potential advantages and drawbacks of OFDM and introduced single-carrier transmission with frequency-domain equalization (SCT-FDE) as an alternative technique. The paper suggested that an SCT-FDE system could achieve the performance of OFDM on frequency-selective multipath radio channels while alleviating its peak-to-average power ratio (PAPR) and synchronization problems. This paper, which was contradicting the claims of many authors, started a long debate, which is still not closed. In the 1994-1995 time period, the same authors published several other papers on the same topic, the most well-known of which being [3].

The OFDM vs. SCT-FDE issue in the 1990s was focused on a pure transmission problem in the context of broadcasting (the wireless communications community was not yet a part of this discussion). In parallel with digital terrestrial television broadcasting, the DVB project was also addressing digital video broadcasting by satellites (DVB-S) and by hybrid fiber/coax (HFC) cable networks (DVB-C). After defining the technical specifications for the broadcast part, the group in charge of the specifications of digital cable TV systems started discussing the return channel for interactive services. One of the proposals was based on a simple orthogonal frequency-division multiple access (OFDMA) system, which assigned one carrier to each subscriber. The carriers were locked to a common source such that the frequency spacing was the inverse of the symbol period used in the transmission. The signals transmitted by the cable modems were therefore single-carrier signals, but the received signal was an OFDM signal. This proposal was rejected by the DVB cable group, but the concept was published in 1996 in [4], which laid the foundation of OFDMA. The word OFDMA itself was coined in this pioneering paper. Several other papers by the same authors followed in 1996-1998, see e.g. [5] and [6].

The motivation for OFDMA in cable TV networks was related to the presence of narrowband interference which affects the uplink. Indeed, TDMA- and CDMA-based systems are very sensitive to this interference and they cannot operate when the interference level exceeds some threshold. In contrast, in an OFDMA system, the cable head-end which assigns resources to cable modems can discard the carriers that are subject to interference and assign only those which have a good signal-to-interference-plus-noise ratio (SINR). The resulting performance improvement over TDMA and CDMA was shown to be substantial [7].

Multicarrier techniques appeared in communications networks with the IEEE 802.11a standard for wireless local area networks and IEEE 802.16-2004 standard for wireless metropolitan area networks. The first of those adopted OFDM for transmission, but multiple access was based on pure TDMA. This kind of OFDM/TDMA was also included in IEEE 802.16-2004, but the standard also featured two other physical (PHY) layers, namely SCT-FDE and OFDMA, and intended to let the market decide. However, the WiMAX Forum, which defines mandatory profiles for fixed WiMAX systems, decided to include the OFDM/TDMA mode only. The IEEE 802.16 group continued its work and released its IEEE 802.16e-2005 specifications for portable and mobile services in 2005. This set of specifications too included 3 PHY layers, but for these applications, the WiMAX Forum selected the OFDMA mode, leading to incompatibility between fixed WiMAX and mobile WiMAX standards.

Another major development in communications networks was born when the Third-Generation Partnership Project (3GPP) started its work to define a technical standard for the so-called Beyond 3G (B3G) systems. Release 8 of the 3GPP standard, which was finalized at the end of 2008, made a large technological gap with respect to previous releases and adopted OFDMA for the downlink and single-carrier FDMA (SC-FDMA) for the uplink. The choice of SC-FDMA for the uplink was motivated by the limited PAPR of this technique
compared to OFDMA.

After this historical review, this paper will give a brief overview of OFDMA in the next section and of SC-FDMA in Section III. Next, in Section IV, we will give some recent performance results comparing the two schemes in a real environment. Finally, we give some conclusions in Section V.

II. BRIEF REVIEW OF OFDMA

Figure 1 presents the baseband structure of a generalized multicarrier (MC) transmitter, which applies to all types of single-carrier (SC) or MC modulation signals transmitted in blocks. Let us denote by \(s_k^{(i)}\) the information symbols (e.g., QAM symbols) which are parsed into data blocks \(x^{(i)}\) of size \(M\). Data blocks belonging to a certain user are precoded with an \(M \times M\) matrix \(M\). The user-specific \(M\)-sized output \(s^{(i)}\) is then mapped onto a set of \(M\) out of \(N\) inputs of the inverse discrete Fourier transform (IDFT) conveniently chosen by the user-specific subcarrier mapping \(N \times M\) matrix \(Q\). \(F_N\) and \(F_N^H\) stand for the \(N\)-point direct and inverse normalized DFT matrices, respectively. A cyclic prefix which needs to be longer than the largest multipath delay is usually inserted before transmission to eliminate the intersymbol interference arising from multipath propagation. In this general representation, OFDMA corresponds to the case without precoding, i.e., the precoding matrix corresponding to OFDMA is the identity matrix \((M = I_M)\). The IDFT operation is equivalent to splitting the information into \(M\) parallel data streams that are transmitted by modulating \(M\) out of the \(N\) distinct subcarriers equally spaced in the channel bandwidth. Thus, OFDMA consists of assigning different subcarrier groups of an OFDM symbol to different users. Compared to an OFDM/TDMA system, which assigns the entire OFDM symbol to one user \((M = N)\), an OFDMA system reduces the granularity in the radio resource allocation mechanism, and this improves the efficiency of the medium access control (MAC) protocol. In addition, an OFDMA system can use the available power more efficiently than a TDMA system. Indeed, focusing on the uplink, an OFDMA system concentrates the power that is available in the user terminal on the carrier group assigned to this terminal, whereas a TDMA system distributes it over the entire channel bandwidth. In an OFDMA systems with \(N\) carriers which allocates \(M\) carriers to each user, the SNR gain with respect to OFDM/TDMA on the uplink is \(10 \log_{10}(N/M)\) dB, which leads to a significant extension of the cell range. Similar gains can also be achieved on the downlink by allocating more power to subcarriers assigned to distant users. But like all other MC schemes, OFDMA suffers from the PAPR problem. Each sample at the IDTF output being the sum of \(M\) independent variables, it is asymptotically Gaussian, and this leads to high envelope variations.

III. SINGLE-CARRIER FDMA

SC-FDMA combines the properties of SC transmission with an OFDMA-like multiple access and attempts to take advantage of the strengths of both techniques: Low PAPR and flexible dynamic frequency allocation.

Depending on the way the subcarriers are allocated to each user or on the way the signal is generated, SC-FDMA can be found in the literature under different names. SC-FDMA was first conceived in a time-domain implementation [8] called IFDMA (Interleaved Frequency Division Multiple Access). At instant \((i)\), blocks of \(M\) data symbols are parsed into data blocks \(x^{(i)}\) of duration \(T = MT_s\), where \(T_s\) is the QAM symbol duration. These blocks are \(K\)-time compressed and 

![Fig. 2. IFDMA signal generation.](image)

As theoretically proven in [9], this manipulation has a direct interpretation in the frequency domain: The spectrum of the compressed and \(K\)-times replicated signal \((F_N X^{(i)})\) has the same shape as the spectrum of the original signal \((F_N X^{(i)})\), with the difference that it includes exactly \(K - 1\) zeros between two data subcarriers, as it can be seen in the example of Figure 3.

This feature enables us to easily interleave a maximum of \(K\) different users in the frequency domain by simply applying to each user a specific frequency shift, or equivalently, by multiplying the time-domain sequence by a user-specific phase ramp. Obviously, this structurally imposes a distributed subcarrier allocation. The spectral considerations above open the way to a frequency-domain implementation of SC-FDMA [10], sometimes called DFT-spread OFDM, and which is in
fact a classical precoded OFDMA scheme, where precoding is done by means of a DFT. This corresponds to using $M = \mathbf{F}_M$ as precoder in Figure 1. Frequency-domain SC-FDMA has a more flexible choice in resource allocation, since matrix $\mathbf{Q}$ can be chosen so as to correspond to contiguous, distributed, mixed or even channel-dependent subcarrier allocation.

The role of the DFT precoder is two-fold: On one hand, this precoding restores the SC-like properties of the signal envelope, alleviating the PAPR problem that is inherent to OFDMA signals. Indeed, we have seen that in the distributed case $y^{(i)}$ is simply the condensed repeated version of $x^{(i)}$, and thus an SC signal. In a localized subcarrier mapping scenario, the spectrum of the SC signal $x^{(i)}$ is simply mapped into a portion of the spectrum of $y^{(i)}$ as in a conventional FDMA system, which does not substantially change the PAPR.

On the other hand, like all precoders, the DFT performs a spreading operation. As a consequence, each modulation symbol $x$ is spread over $M$ subcarriers. This introduces some built-in frequency diversity, and losing the information on one subcarrier because of a fading dip does not lead to losing all the information in a modulation symbol as in OFDMA. But spreading does not only have beneficial consequences. It also causes intercode interference on frequency selective channels. Indeed, frequency selective fading causes a loss of orthogonality between the $M$-sized spreading codes. This affects all the modulation symbols composing $x^{(i)}$, and the effect is especially disturbing with high-order modulations, as it will be shown in the simulations section.

### IV. PERFORMANCE RESULTS

In this section, we report some simulation results obtained for the uplink of the LTE systems. Among $N = 512$ subcarriers which compose the transmitted signal, 300 are modulated data carriers, the remaining 212 being reserved as guard bands. The 300 data carriers are split into 25 resource blocks (RBs) of $M = 12$ subcarriers. After data scrambling, we use a turbo code (TC) with different rates prior to QAM signal mapping. A cyclic prefix with a length of 31 samples is employed. Groups of 12 SC-FDMA symbols are encoded together and sent through a vehicular $A$ channel profile with 6 taps and a maximum delay spread of 2.51 $\mu$s. Perfect channel state information (CSI) was assumed in a first step. The channel bandwidth was 5 MHz and the sampling frequency was 7.68 MHz. The results are reported in Figures 4, 5, and 6, for QPSK, 16-QAM, and 64-QAM, respectively. In the simulations, 5 localized RBs (60 localized subcarriers) are allocated to each user.

Figure 4 shows the results for QPSK. Since OFDMA has no built-in diversity, its performance is very dependent on the coding rate. When a high coding rate is employed or the system is uncoded, OFDMA performs poorly because coding does not manage to compensate the influence of subcarriers with a low SNR. When stronger coding is present (e.g., rate 1/2), OFDMA benefits from the coding diversity and recovers its performance loss and even outperforms SC-FDMA by 0.5 dB at the frame error rate (FER) of 1%.

With higher-level modulations, there is a tradeoff between the frequency diversity gain (due to the spreading performed in SC-FDMA) and the intercode interference caused by the frequency selectivity of the channel. This tradeoff is also driven by the coding rate. Let us examine the FER results in Figures 5 and 6, centered on a target FER of 1%. We notice that SC-FDMA is more sensitive to intercode interference when the modulation order increases (16-QAM, 64-QAM). In this case, coded OFDMA has better performance. The higher the modulation order, the stronger this effect is: OFDMA with code rate 1/2, for example, outperforms SC-FDMA by 0.6 dB, 2.6 dB and 4.4 dB when employing QPSK, 16-QAM and 64-QAM respectively. SC-FDMA outperforms OFDMA when low modulation order (QPSK) or uncoded modulation is employed. The results are summarized in Table I.
In Figure 4, distributed vs. localized subcarrier allocations are also investigated. Localized subcarrier mapping has poorer performance as it provides less frequency diversity than distributed mapping. Nevertheless, in practice, the channel estimation errors are more important in the distributed subcarrier mapping: Here, Wiener filtering was used to estimate the channel. In a localized subcarrier mapping scenario, we can take advantage of the channel’s correlation profile in the frequency domain in order to maximize the SNR of the estimation, while in distributed subcarrier mapping this is not possible since the pilots experience uncorrelated channel realizations. This problem becomes even more critical in the case of multiple transmit antennas. The FER performance advantage due to higher frequency diversity in the distributed case is lost because of channel estimation difficulties, and localized scenarios are preferred in LTE uplink. To gain some frequency diversity, localized subcarrier mapping with frequency hopping (FH) is an interesting option, especially in the case of small spectral allocations.

Next, we investigated the impact of high-power amplifier (HPA) nonlinearities on the performance of these techniques using the requirements of the LTE system. Three main requirements need to be fulfilled here.

These are: (1) the spectrum emission mask (SEM), (2) the in-band distortion (which is measured in percentage by the error vector magnitude, EVM), and (3) the out-of-band emissions (limited by the adjacent channel leakage ratio, ACLR). Numerical bounds for these constraints are given in [11]. ACLR and EVM are respectively bounded by a minimum attenuation of 30 dB for the out-of-band emissions and maximum in-band signal distortion of 17.5% in the case of QPSK.

To make good use of the available power, it is necessary to operate the HPA near saturation. But this results in nonlinear signal distortion, which is higher in the case of signals of high dynamic range. To reduce signal distortion, the HPA output power needs to be backed-off with respect to its saturation level. High output back-off (OBO) values reduce distortion, but they also reduce the power efficiency. To optimize the operating point of the HPA in terms of output power and signal distortion, we need to minimize the total SNR degradation which is defined as the sum of the OBO and of the resulting SNR degradation caused by nonlinear signal distortion. The impact of the nonlinearities and the different components of the nonlinear degradation suffered by a system are summarized in Figure 7.

As illustrated in Figure 7, there exists an optimum operating point $I_{\text{opt}}$ (and thus an optimum value of the OBO) which minimizes the total degradation. Unfortunately, $I_{\text{opt}}$ is not always possible to achieve in practical systems due to spectral mask requirements. The optimum operating point $I_{\text{opt}}$ in this figure corresponds to an OBO of 4 dB and leads to a total in-band degradation of approximately 5.4 dB. But, $I_{\text{opt}}$ lies in a region which corresponds to high levels of out-of-band emissions, and might also cause high EVM. Therefore, the HPA in this figure must operate at a point $I$ which is the closest point to $I_{\text{opt}}$ which satisfies all system requirements (ACLR, SEM, EVM). The gap between the operating point $I$ and the optimum point $I_{\text{opt}}$ may attain several dB in real systems.

With a precoded-OFDMA system as described at the beginning of this section, we used the Rapp model with knee factor 2 [12] as well as the Saleh model with $\alpha = 1, \beta = 1/4, \alpha_p = \beta_p = 1$ [13] for the HPA nonlinearity and evaluated the optimum back-off for the amplifier under system constraints. The simulations were carried out for localized subcarrier allocations and for different numbers of RB allocations to users, with QPSK signal mapping, as shown in Table II. Detailed results including distributed carrier mapping are given in [14] and [15]. The Rapp amplifier model exhibits amplitude distortion, but no phase distortion. The in-band distortion (measured by EVM levels) is less significant than in the case of Saleh model, also introducing phase distortions.

With the Rapp model, the SEM is the strongest constraint, while with the Saleh model EVM is the strongest constraint and operating points lie at much higher values of the back-off due to the more pronounced nonlinear HPA characteristics.

Due to its PAPR advantage, SC-FDMA systematically gains 1.5 - 2 dB in terms of OBO, thus offsetting its performance loss on wireless channels. This is confirmed in Figure 8, where comparative total degradation results adding up the effects of both nonlinearities and behavior in frequency selective channels are presented. SC-FDMA, which was outperformed
CIOCHINA AND SARI: A REVIEW OF OFDMA AND SINGLE-CARRIER FDMA AND SOME RECENT RESULTS

TABLE II
COMPARATIVE PERFORMANCE OF OFDMA AND SC-FDMA UNDER SYSTEM CONSTRAINTS

<table>
<thead>
<tr>
<th>Under SEM constraints</th>
<th>SC-FDMA</th>
<th>OFDMA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 RB</td>
<td>5 RB</td>
</tr>
<tr>
<td>Rapp HPA</td>
<td>OBO (dB)</td>
<td>4.5</td>
</tr>
<tr>
<td>EVM (%)</td>
<td>14.6</td>
<td>11.6</td>
</tr>
<tr>
<td>ACLR (dB)</td>
<td>30.9</td>
<td>31.7</td>
</tr>
<tr>
<td>OBO (dB)</td>
<td>8.9</td>
<td>8.9</td>
</tr>
<tr>
<td>Saleh HPA</td>
<td>EVM (%)</td>
<td>17.4</td>
</tr>
<tr>
<td>ACLR (dB)</td>
<td>31.6</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Fig. 7. Illustration of nonlinear distortion in an uncoded OFDMA system with 16-QAM signal mapping.

Fig. 8. Total system degradation of OFDMA and SC-FDMA, QPSK 1/2, 5 localized RBs, target FER 1%, Rapp HPA.

by OFDMA by 0.5 dB on the wireless channel, turns out to have an OBO advantage of 2 dB due to its better PAPR performance. Overall, the gain of SC-FDMA over OFDMA in this case amounts to 1.5 dB. Note that the part of the SNR loss in the total degradation balance (visible from the convex shape of the total degradation curve corresponding to low OBO in Figure 7) is highly reduced in coded systems. Indeed, the in-band distortion is mostly eliminated by the error correcting code when operating in the range of reasonable OBO values. The operating point is pushed into the linear part of the total degradation curve in Figure 8, because the SNR degradation due to in-band distortion (on the order of 0.2 dB) is completely negligible with respect to OBO values, which are on the order of several dB.

On the other hand, when the modulation order increases in the presence of strong coding, OFDMA becomes more and more attractive, since two effects combine: The potential OBO gain decreases (less PAPR difference), and the performance gain of OFDMA over SC-FDMA strongly increases, as it has been shown in Figures 4 and 5.

V. CONCLUSIONS

In this paper, we have given a historical review of two popular multiple access techniques, which are OFDMA and SC-FDMA. The controversial SCT-FDE vs. OFDM issue, which started in the early 1990s at the time of the European DVB project, continues today as an SC-FDMA vs. OFDMA debate in wireless communications. Whereas OFDMA was selected by the WiMAX Forum for mobile WiMAX systems for both downlink and uplink, the 3GPP project preferred to use OFDMA for the downlink only and favored SC-FDMA for the uplink.

We have also reported the results of some recent work on performance evaluation of these two multiple access techniques, which indicate that both techniques have some virtues and neither of them is better than the other in all conditions. In summary, OFDMA turns out to have better performance with high-order modulations which are used in favorable propagation conditions (typically for users near the base station). Stated differently, OFDMA lowers the SNR threshold above which high-level modulations and high code rates can be used. In contrast, SC-OFDMA is superior with QPSK and low code rates used typically near the cell edge and for users with bad propagation conditions. As a result, OFDMA can be expected to offer a higher cell capacity, while SC-FDMA can lead to cell range extension.

REFERENCES


Cristina Ciochina was born in Bucharest, Romania, in 1980. She received her Engineering degree in Electronics and Telecommunications from the Polytechnic University of Bucharest, Romania, in 2004. She received her M.Sc. degree from the Ecole Supérieure d’Electricité (SUPELEC), France, in 2005 and her Ph.D. degree from the Université de Paris-Sud 11, France in 2009, both in Telecommunications.

Since 2005, she has been with Mitsubishi Electric R&D Centre Europe (MERCE), France. Her technical interests include physical layer design, multicarrier systems, signal processing for future wireless communications systems, and 3GPP/LTE. Dr. Ciochina has published 18 technical papers and is the author or co-author of 8 patent applications.

Hikmet Sari (S’78 - M’81 - SM’88 - F’95) received his Diploma (M.S.) and Ph.D. in Telecommunications Engineering from the ENST, Paris, France, in 1978 and 1980, respectively, and the Habilitation degree from the University of Paris-Sud, Orsay in 1992.

He was with Philips Research Laboratories from 1978 to 1989, first as Researcher and then as Group Supervisor. From 1989 to 1996, he was R&D Department Manager at SAT (SAGEM Group), and from 1996 to 2000, he was Technical Director at Alcatel. In May 2000, he became Chief Scientist of the newly-founded Pacific Broadcom Communications, which was acquired by Juniper Networks in December 2001. Since April 2003, he has been a Professor and Head of the Telecommunications Department at SUPELEC, and since December 2004 he is also Chief Scientist of Sequans Communications.

Dr. Sari was an Editor of the *IEEE Transactions on Communications* from 1987 to 1991, a Guest Editor of the *European Transactions on Telecommunications (ETT)* in 1993, a Guest Editor of the *IEEE JSAC* in 1999, an Associate Editor of the *IEEE Communications Letters* from 1999 to 2002, and a Guest Editor of the EURASIP Journal on Wireless Communications and Networking in 2007. He was also Chair of the Communication Theory Symposium of the 2002 IEEE International Conference on Communications (ICC 2002), April 2002, New York, Technical Program Chair of ICC 2004, June 2004, Paris, and Vice General Chair of ICC 2006, June 2006, Istanbul. He is currently serving as General Chair for the forthcoming 2010 Annual IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2010) and the 2012 IEEE Wireless Communications and Networking Conference (WCNC 2012). He was elevated to the IEEE Fellow Grade and received the Andre Blondel Medal from the SEE (France) in 1995 and he received the Edwin H. Armstrong Achievement Award from the IEEE Communications Society in 2003.
Bayesian Foundations of Channel Estimation for Smart Radios

Romain Couillet, Andrea Ancora, and Mériouane Debbah

Abstract—In this paper, we revisit the philosophical foundations of the field of channel estimation. Our main intention is to come up with a partial answer to the question: “given some available sensed signals, how should cognitive radios ideally perform channel estimation?” We specifically introduce a general framework to provide optimal channel estimates under any prior knowledge at the sensing device. Our discussion is articulated as a top-down approach, introducing successively (i) a discussion on the philosophical foundations of channel estimation as a simplification measure for the general problem of wireless detection, (ii) an information theoretically optimal approach to channel detection assuming the sensing device has infinite memory, and (iii) a derived optimal approach when limited memory size is accounted for. The key mathematical tools used in this discussion emerge from Bayesian probability theory and are known as the maximum entropy principle and the minimum update principle. Derivations are carried out for the particular case of channel estimation in orthogonal frequency division multiplexing (OFDM) systems. While some theoretical results will be proven to match already known techniques, such as Kalman filters, another set of novel results will be shown by simulations to perform better than known channel estimation schemes.

I. INTRODUCTION

Channel estimation, along with most synchronization procedures, is itself a major and historical field of research in the realm of wireless communications. As such, thousands of novel channel estimators are proposed and compared to previous estimators every year. One of the obvious reasons for such an activity around channel estimation is that there does not exist a universal measure of performance to rate any given scheme with respect to any other; instead, several different selection criteria are considered, such as computational complexity, mean square error of the estimate, robustness against outage channel conditions, required processing memory etc. In this work, we wish to propose a framework to encompass the aforementioned selection criteria into a unique channel estimation framework. In some cases, we shall generate the optimal channel estimators for this framework. The sought for a general framework for channel estimation is motivated by the recent trend towards cognitive radios, and in particular by the trend towards developing terminal-centered intelligence in future wireless flexible networks. In the framework of cognitive radios, terminals are required to learn from their environment and take optimal decisions based on a constantly updated knowledge: performing optimal channel estimation from limited side information is part of the requirements demanded of cognitive radios.

The prior requirement for channel estimators is to help signal decoders as best as possible. The decision to consider a particular measure of performance for a given channel estimators should only reflect the subsequent effects on the eventual signal decoding process. For instance, minimizing the mean square error of channel estimators has no theoretical ground; it is merely a convenient mathematical way to compare estimators. In Section II, we shall briefly remind the foundations of signal decoding and channel estimation, which we shall review on a cognitive radio viewpoint. This study naturally follows the theoretical ideas introduced in [5], in which the authors introduce a new look on cognitive radios, and in [6], where the specific problem of blind source detection is addressed. Section II will conclude that today’s theoretical and technological advances does not yet allow smart devices to perform optimal signal decoding without considering channel estimation as an independent entity.

In Section IV, we shall therefore treat channel estimation as a self-contained process, independent of the problem of signal detection, as is conventionally the case. We shall introduce a complete framework to derive optimal channel estimators under any prior state of knowledge at the signal receiver. While the conventional approach is to treat specific channel models and develop estimators for those models, we shall here instead consider prior knowledge about the environment at the receiver, and develop consistent estimators for this knowledge. We wish indeed to insist on the fact that smart devices should be able to come up with an ideal channel estimator for any given prior information on the channel. The approach addressed here is based on conventional Bayesian probability theory and on the maximum entropy principle [8]. In this section, we shall essentially remind the results originally derived in [7]. However, the practical finite memory size of the processing devices will not be taken into account in this section.

To answer the problem of optimal channel estimation under finite memory-size constraint, we shall subsequently introduce a novel aspect of the maximum entropy principle, known as the minimal update principle [9]. The major difference between both is the fact that the maximum entropy principle assumes an initial starting point and performs optimal decisions for data collected from this starting point on, while the minimal update principle assumes that one might be oblivious of old data and performs optimal decisions for a finite window of “remembrance”. In our specific channel estimation considerations, this means that estimation based on the maximum

Romain Couillet is with ST-Ericsson, Supélec, 635 Route des Lucioles, 06560 Sophia Antipolis, France (email: romain.couillet@superlec.fr)
Andrea Ancora is with ST-Ericsson, Eurécom, 635 Route des Lucioles, 06560 Sophia Antipolis, France (email: andrea.ancora@eurecom.fr)
Mériouane Debbah is with Alcatel-Lucent Chair, Supélec, Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif sur Yvette, France (email: meriouane.debbah@superlec.fr)
entropy principle assume infinite memory at the receiving side information is available is treated by Jaynes, through the maximum entropy principle [8]. However MaxEnt does not allow to perform any estimation of the channel with only a priori knowledge of the transmitted signal.

Indeed, while Shannon [10] allows us to derive the capacity of a channel for which all synchronization parameters, plus the noise variance, are perfectly known, no such theory exists when the knowledge of some of these parameters is missing. More precisely, for a scalar communication channel for which all synchronization parameters, plus the noise variance, are perfectly known, no such theory exists, and the maximum likelihood estimator for Gaussian $n$ and uninformative $I$ becomes

$$
\hat{x} = \arg \min_{x \in \mathcal{X}} \| y - h_x \|^2 \quad (8)
$$

which requires to have an a priori $p(h)$ for $h$. But this a priori is too impractical to obtain and would require to know all possible channel realizations and their respective probability. As a consequence to this strong difficulty, most contributions in the synchronization field have provided various empirical models based on field observations in order either to give an expression to $p(h)$ or, more practically, to propose good channel estimators $\hat{h}$ to $h$. Among those solutions, we mention [1] [2] [3] [4].

The difficulty of handling estimation problems when little side information is available is treated by Jaynes, through the Bayesian probability field, thanks to the maximum entropy principle [8]. However MaxEnt does not allow to perform updates of probability when new information, such as new pilots in the channel estimation problem, is available. In this case, the complete set of past symbols plus a priori distribution for the channel $h_0$ at time $t = 0$. This question was treated in [12] for the OFDM framework, when the channel delay spread, the channel time correlation and the signal-to-noise ratio (SNR) are alternatively known or unknown. When these parameters are not perfectly known, MaxEnt provides channel estimates minimizing the estimate mean square error (MMSE estimates) that outperform classical estimates which use empirical (often erroneous) models. Recent contributions in the Bayesian probability field enable one to perform probability updates, in particular based minimum cross entropy considerations [9]. In this work, we will then provide a channel estimation method, using the ME principle, which allows to assign probability distributions for the channel, when the estimator only knows the last past inferred channel distribution and the new received pilot symbols. The remainder of this article unfolds as follows: in Section II, we discuss the foundations of channel estimation under a Bayesian point of view; in Section III, we introduce the OFDM model that shall be used as a toy example to illustrate in practice the theoretical ideas elaborated in the following sections. In Section IV, we discuss optimal infinite memory channel estimation, while in Section V, we introduce the minimal update principle and extend the previous optimality framework to finite-time memory channel estimation; also in this section, technical comparison is made against classical techniques. In Section VI, simulation and results are proposed, which compare the new method to the aforementioned classical algorithms. Finally, in Section VII, we draw our conclusions.

Notations: In the following, boldface lower case symbols represent vectors, capital boldface characters denote matrices ($I_N$ is the $N \times N$ identity matrix). The transposition operation is denoted $(\cdot)^T$. The Hermitian transpose is denoted $(\cdot)^H$. The operator $\operatorname{diag}(\mathbf{x})$ turns the vector $\mathbf{x}$ into a diagonal matrix. The symbol $\det(\mathbf{X})$ is the determinant of matrix $\mathbf{X}$. The symbol $\mathbb{E}[\cdot]$ denotes expectation. The Kronecker delta function is denoted $\delta_{x}$ that equals 1 if $x = 0$ and equals 0 otherwise.

II. FOUNDATIONS OF CHANNEL ESTIMATION

In 1948, Shannon [10] provided the expression for the capacity of a communication channel between a transmitter and a receiver, modelled as

$$
y = x + n \quad (2)
$$

where $x \in \mathcal{X}$ is the input signal sent by the transmitter, which the receiver aims at recovering, $n$ some additive noise process and $y$ the effective signal captured by the receiver. In the conventional case when $n$ and $x$ are random variables with zero mean Gaussian distributions, the rate $C$ to which the sequence of $x$ can be decoded with infinitely low decoding error takes the simple form

$$
C = \log \left( 1 + \frac{\mathbb{E}[|x|^2]}{\mathbb{E}[|n|^2]} \right) \quad (3)
$$

but Shannon does not provide any way to achieve such a decoding rate. Although, for appropriate coding schemes, it is possible to get an estimate $\hat{x}$ of $x$ with as low decoding error rate as desired. The estimate $\hat{x}$ is based on the posterior probability $p(x|y)$ of any candidate $x \in \mathcal{X}$ given the output $y$,

$$
p(x|y, I) = \int_n p(x|y, n, I)p(n|I)dn \quad (4)
$$

where we denote by $I$ all prior information known by the receiver at the moment it receives $y$.

In particular, we often take $\hat{x}$ to be the maximum likelihood estimator for $x$, and

$$
\hat{x} = \arg \max_{x \in \mathcal{X}} \int_n p(x|y, n, I)p(n|I)dn \quad (5)
$$

which is easily derived for Gaussian $n$ and $I$ bringing no information to $n$, as

$$
\hat{x} = \arg \min_{x \in \mathcal{X}} \| y - x \|^2 \quad (6)
$$

When the signal $x$ is filtered by a channel $h$, i.e. $y = hx + n$, the previous derivation is still valid, and we get the posterior probability $p(x|I)$ as,

$$
p(x|y, I) = \int_h \int_n p(x|y, n, h, I)p(n|I)p(h|I)dn\,dh \quad (7)
$$

If $h$ is known, this boils down to a scaled version of the previous scenario, and the maximum likelihood estimator for Gaussian $n$ and uninformative $I$ becomes

$$
\hat{x} = \arg \min_{x \in \mathcal{X}} \| y - hx \|^2 \quad (8)
$$
However, one rarely has access to the exact value for $h$ and then the true maximum likelihood estimate for $x$ is simply

$$
\hat{x} = \arg \max_{x \in \mathcal{X}} \int_h \int_n p(x|y, n, h, I)p(n|I)p(h|I)dh \, dn
$$

When one performs channel estimation, one gets some estimate $\hat{h}$ of the true channel $h$ from previously received pilots, gathered in the information $I$. Classically, this estimate is then directly used in (9) by replacing the term $p(h|I)$ by $\delta(h - \hat{h})$. This substitution however constitutes a major mathematical flaw, unless some (possibly malevolent) genie ensured the receiver that the true channel is $\hat{h}$ with probability one. As such, the whole field of channel estimation has no information theoretical grounds. However, solving (9) in general is an extremely involved problem, which requires integration over all possible $h$ channels. Note by the way that $h$ might be a multi-dimensional vector channel, so that the integration over $h$ might in truth be a multi-variate integral. It seems therefore natural to approximate the integration (9) by replacing the term $p(h|I)$ by $\hat{h}$ for a finite set of $n$ candidates $h_1, \ldots, h_n$ with high probability.

The question of the choice of $n$ and $h_1, \ldots, h_n$ is a rather involved problem, which, to the authors' knowledge has not yet been addressed. This is however not the purpose of the current work. Instead, we shall focus on the case $n=1$, where $h_1$ is an estimate of the true channel $h$, based on coherent information provided from pilots and previous data, all captured in $I$. The question that now arises is: what measure should the channel estimator minimize? What are the grounds for performing minimum mean square error (MMSE), maximum likelihood (ML) estimation? The correct estimator for $h$ should be that estimator $\hat{h}$, which is such that the posterior probability $p(x|y, h, I)$ when $h$ is known is “close” to $p(x|y, \hat{h}, I)$ when $h$ is unknown. Taking the classical Kullback-Liebler distance to compare distances between probability distributions, we may then consider

$$
\tilde{h} = \arg \min_{h_1} p(x|y, h_1, I) \log \frac{p(x|y, h, I)}{p(x|y, h_1, I)}
$$

The computational difficulty of the above expression however leads one to consider more tractable channel estimate minimization functionals, such as MMSE or ML channel estimators. The other difficulty arises from a consistent evaluation of $p(h|I)$ when $I$ encompasses pilots, prior data and overall prior information about the channel at the receiver. It is indeed rather intricate to provide a mathematically sound description of $I$. The purpose of the subsequent sections will be to cast some light on a general Bayesian framework to evaluate $p(h|I)$, applied to the concrete case of channel estimators for OFDM systems based on pilots. The next section is dedicated to introducing the model for OFDM transmission channels.

Fig. 1. Time-frequency OFDM grid with pilot positions enhanced

III. CASE STUDY: OFDM CHANNEL MODEL

Consider a single cell OFDM system with $N$ subcarriers. The cyclic prefix (CP) length is $N_{CP}$ samples. In the time-frequency OFDM symbol grid, pilots are found in the symbol positions indexed by the function $\phi_k(n) \in \{0, 1\}$ which equals 1 if a pilot symbol is present at subcarrier $n$, at symbol time index $t$, and 0 otherwise. We further denote $P_t \in \mathbb{R}^{N_{CP} \times N}$ a diagonal matrix with $(i, i)$ entry $P_{t,ii} = \phi_k(i)$. The time-frequency grid is depicted in Figure 1. Both data and pilots at time $t$ are modeled by the frequency-domain vector $\mathbf{s}_t \in \mathbb{C}^N$ with pilot entries of zero mean and amplitude $|s_{t,k}|^2 = 1$. The transmission channel is denoted $\mathbf{h}_t \in \mathbb{C}^{N_{CP}}$ in the frequency-domain and is known only to have overall power 1. The additive noise is denoted $\mathbf{n}_t \in \mathbb{C}^{N_{CP}}$ with entries known to have total variance $\sigma^2$. The time-domain representation of $\mathbf{h}_t$ is denoted $\mathbf{v}_t \in \mathbb{C}^L$ with $L$ the channel length, i.e. the channel delay spread expressed in OFDM-sample unit. The frequency-domain received signal $\mathbf{y}_t \in \mathbb{C}^{N_{CP}}$ is then

$$
y_t = \text{diag}((\mathbf{h}_t)s_t + \mathbf{n}_t)
$$

We will also denote, $\forall k \in \{1, \ldots, N\}$, $h'_k = y_k/s_k = h_k + n_k/s_k$ and $h' = (h'_1, \ldots, h'_N)^T$ (here, the time index $t$ is implicit).

The channel $\mathbf{h}_t$ evolves in time with coherence time function $\lambda(\tau)$ such that, independently of the channel delay spread index

$$
E[\nu_{t,\lambda(\tau)}] = \frac{\lambda(\tau)}{L}
$$

Along this study, we might consider the different system parameters, such as $\lambda(\tau)$ to be either exactly known at the receiver (and then fully part of the prior information $I$) or only partially known. In the following section, we establish, under partial or total knowledge of the different system parameters, an optimal framework for channel estimation. For the OFDM example, this section mainly recalls the results of [7].
IV. MAXIMUM ENTROPY CHANNEL ESTIMATION

The essential derivations of this section will consist in establishing, at time $t$, the posterior probability

$$p(h_t | y_t, y_{t-1}, \ldots, y_1, I)$$

(14)

For readability we only treat the case $t = 2$, the general case being a trivial extension. Assume only pilots are transmitted or, at least, that the information about $I$ carried by the non-pilot signals are rather uninformative. Discarding the terms $I$ for readability, we have in that case

$$p(h_2 | y_2, y_1, I) = p(h_2 | h'_2, h'_1, I)$$

$$= p(h_2 | h'_2) p(h'_1 | h_2) / p(h'_1 h'_2)$$

$$= \frac{p(h_2) p(h'_2 | h_2) p(h'_1 | h_2) d h_1}{p(h'_1 h'_2)}$$

(15)

(16)

(17)

When the exact time evolution model for $h_t$ is known, $p(h_1 | h_2)$ has an explicit expression which allows then to perform the above calculus. In practice however, it is rarely the case that such a time-evolution model be perfectly known. One then needs here an automatic method to provide a consistent expression for $p(h_1 | h_2, I)$ when little is known about the interaction between $h_1$ and $h_2$. The method we shall use here is referred to as the maximum entropy principle [8].

Consider a given parameter $x$, whose knowledge is limited to the information contained in $I$. The maximum entropy principle allows one to assign a unique probability distribution $p(x | I)$ as follows,

1) among the set of all probability distributions, consider those distributions that satisfy the constraints about $x$ given in $I$. This is, we exclude all distributions that do not satisfy the prior information $I$. The remaining set of probability distributions is denoted $\Omega$.

2) in this remaining set of acceptable distributions $\Omega$, $p(x | I)$ is assign the distribution which has maximum entropy, i.e.

$$p(x | I) = \arg \max_{q \in \Omega} \int q(y) \log q(y) dy$$

(18)

Choosing the distribution that has maximum entropy allows one not to make undesired assumptions on the unknown system variables, and as such allows one to remain neutral with regard to unaccessible information; see [11] for further details on the maximum entropy principle. In the model described in Section III, since only the total power is known about both the channel delay profile and the additive noise, both are assigned Gaussian distributions with zero mean and variance consistent with prior knowledge, as required by the maximum entropy principle. We then have $n_t \sim \mathcal{CN}(0, \sigma^2 I_N)$ and $n_u \sim \mathcal{CN}(0, \frac{1}{L} I_L)$, which in the frequency domain, applying Fourier transform, translates into $h_t \sim \mathcal{CN}(0, Q)$, with $Q$ defined as

$$Q_{nm} = E \left[ \sum_{k=0}^{L-1} \sum_{l=0}^{L-1} \nu_k \nu_l^* e^{-2\pi i k n / N} \right] = \frac{1}{T} \sum_{k=0}^{L-1} e^{-2\pi i k m / N}$$

(19)

Note that $Q$ is singular, since $L < N$ as OFDM requires.

If the coherence time function $\lambda(t)$ is perfectly known, the maximum entropy principle then assign to $p(h_t | h_{t+r})$ a Gaussian distribution of mean $\lambda(t) h_{t+r}$ and variance

$$E[h_t, h_{t+r}] = (1 - \lambda(t)^2) Q$$

(20)

In the problem with $t = 2$, we denote $\lambda = \lambda(1)$. We then find that $p(h_1 | h_2, I)$ is Gaussian and satisfies

$$p(h_1 | h_2) = \lim_{t \to \phi} \frac{1}{\pi N_0} e^{-\lambda h_1^2}$$

(21)

for $\{\phi\}$ any sequence converging to $\Phi$. This allows then to compute the full expression of $p(h_2 | y_2, y_1, I)$ and a closed-form expression of some conventional estimators. In particular, the MMSE estimator $\hat{h}_2^{(\text{MMSE})}$ for $h_2$, defined as

$$\hat{h}_2^{(\text{MMSE})} = E[h_2 | y]$$

(22)

in this case expresses as [7]

$$\hat{h}_2^{(\text{MMSE})} = M_2^{-1} \left( \begin{bmatrix} P_L \nu_2 \nu_1^* & (I_N + \lambda^2 \nu_2 \nu_1^*)^{-1} \lambda \nu_1 \nu_1^* \end{bmatrix} \right)$$

(23)

with $M_2$ satisfying

$$Q M_2 = \frac{I_N}{1 - \lambda^2} - \frac{\lambda^2}{1 - \lambda^2} (I_N + \lambda^2 \nu_2 \nu_1^*)^{-1} + \frac{\nu_2 \nu_2^*}{\sigma^2}$$

(24)

This expression generalizes to $t \geq 1$, for which we have the MMSE estimator $\hat{h}_2^{(\text{MMSE})}$ given in Equation (25).

However it often occurs that the assumption that $L$ and $\lambda(1), \lambda(2), \ldots$ are known a priori at the receiver is not realistic. In truth, rather limited information is known a priori on these parameters. We may then reconsider (17) to include the uncertainty on $L$ and/or $\lambda(1), \lambda(2), \ldots$. In the previous case $t = 2$ setup, we have in particular

$$\hat{h}_2^{(\text{MMSE})} = E[h_2 | y_2, y_1, I]$$

(26)

$$= \int p(\lambda | I) p(L | I) E[h_2 | y_2, y_1, \lambda, L, I] d \lambda d L$$

(27)

which leads to yet other expressions derived thoroughly in [7].

Through the OFDM example, we therefore developed a rather automatic method to derive consistent estimators under any prior knowledge at the receiver which we claim optimal on an information theoretic viewpoint, i.e. those estimators are taking into consideration all prior information $I$ and are made such that no ad-hoc assumption is taken regarding imperfectly known parameters, while being compliant with the Bayesian principles.

However, an underlying assumption of the previous approach is that infinite storage is available at the receiver. Indeed, for large $t$, we still need to consider all events from time instants $1$ to $t$; if we decide to discard the oldest data, we then depart from the optimality of the proposed scheme. We then need to reconsider the whole framework to include an additional feature: the receiver is oblivious to part of the past events. The natural way to handle this modification of the current maximum entropy framework is to consider updated distribution assignments for posterior probabilities instead of absolute distribution assignments. This is the topic of the next and our main section.
\[ \hat{h}(\text{MMSE}) = \left( 1 + \sum_{k=1}^{t} \frac{\lambda(k)^2}{1 - \lambda(k)} \right) I_N - \sum_{k=1}^{t} \frac{\lambda(k)^2}{1 - \lambda(k)} \left( I_N + \frac{1 - \lambda(k)^2}{\sigma^2} Q \right)^{-1} \]  

\[ \times \left( \sum_{k=1}^{t} \lambda(k) \left( I_N + \frac{1 - \lambda(k)^2}{\sigma^2} P_k Q \right) + I \right)^{-1} \frac{1}{\sigma^2} P_k \hat{h}_k' \]  

(25)

V. Minimal Channel Estimation Update

A. Introduction to Bayesian minimal update

When it comes to update probability assignments, Caticha proposes an extension of the maximum entropy principle, namely the minimum cross entropy principle (ME) [9]. When \( p(h_t | I_1) \) has been assigned for some side information \( I_1 \), and new cogent information \( I_2 \) is later available, then the ME principle consists in assigning to \( p(h_t | I_2) \) the distribution

\[ p(h_t | I_2) = \arg \min_q S[q, p(h_t | I_1)] \]  

(28)

where

\[ S[q, p] = \int q(x) \log \left( \frac{p(x)}{q(x)} \right) dx \]  

(29)

The functional \( S[q, p] \) is referred to as the cross-entropy between the probability distributions \( q \) and \( p \).

This method is based on a minimal update requirement, which in essence assigns to \( p(h_t | I_2) \) the unique distribution which minimizes the changes brought to \( p(h_t | I_1) \) while satisfying the new constraints given by \( I_2 \). In the following, additional side information on \( h_t \) (which possibly varies over time) will come from new available pilots at later time positions.

B. Perfect system parameters knowledge

We assume here that channel estimation is performed at different time instants \( t = 1, 2, \ldots \). Denote \( I_k \) the knowledge at time \( k \). Since memory restrictions impose to discard past received data, we decide here only to consider at time \( k \) the last received pilot data symbols, the last assigned probability \( p(h_k | I_{k-1}) \), and the supposedly known time correlation \( \lambda = \lambda(1) \) between the current channel \( h_k \) and the past channel \( h_{k-1} \).

Assume prior assigned distribution \( p(h_{k-1} | I_{k-1}) \) at time index \( k-1 \). We have in general

\[ p(h_k | y_k, I_k) = \int p(y_k | h_k, I_k) \cdot p(h_k | I_k) \]  

(30)

\[ = p(y_k | h_k, I_k) \cdot \int p(h_k | h_{k-1}, I_k) \cdot p(h_{k-1} | I_k) dh_{k-1} \]  

(31)

\[ = \frac{p(y_k | h_k, I_k)}{p(y_k | I_k)} \]  

(32)

Let us now perform a recursive reasoning over the channel estimates at time indexes \( k \in \mathbb{N} \). Assume that \( p(h_{k-1} | y_k, I_{k-1}) \) is Gaussian \( \mathbb{C} \mathbb{N}(k_{k-1}, M_{k-1}) \). We will show that this implies \( p(h_k | y_k, I_k) \) is still Gaussian. This will therefore be denoted \( \mathbb{C} \mathbb{N}(k_k, M_k) \). We have

\[ p(h_k | y_k, I_k) = \alpha_t p(h_k | I_k) \int p(h_k | h_{k-1}, I_k) \cdot p(h_{k-1}, I_k) dh_{k-1} \]  

(33)

\[ = \lim_{Q \rightarrow Q} e^{(h_k - \lambda h_{k-1})^H Q^{-1} (h_k - \lambda h_{k-1})} \int e^{(h_{k-1} - \lambda h_{k-1})^H Q^{-1} (h_{k-1} - \lambda h_{k-1})} \cdot \alpha_t (h_{k-1} - \lambda h_{k-1})^H M_{k-1}^{-1} (h_{k-1} - \lambda h_{k-1}) dh_{k-1} \]  

(34)

\[ = (h_k - \lambda h_{k-1})^H Q^{-1} h_k - \lambda h_{k-1} \]  

(35)

where the \( \dot{Q} \)'s are taken from a set of invertible matrices in the neighborhood of \( Q \), and the \( \alpha_t \)'s are constants.

First we need to write the products in the integrand in a single Gaussian exponent form

\[ \frac{1}{2} \int_{-\infty}^{\infty} \dot{Q}^{-1} dh_k \]  

(36)

\[ \text{with} \begin{cases} N = \frac{\lambda^2 \dot{Q}^{-1}}{1 - \lambda^2} + M_{k-1}^{-1} \\ L = N^{-1} \left( \frac{\lambda^2 \dot{Q}^{-1}}{1 - \lambda^2} h_k + M_{k-1}^{-1} k_{k-1} \right) \\ C = h_k^H \frac{\lambda^2 \dot{Q}^{-1}}{1 - \lambda^2} h_k + k_{k-1}^H M_{k-1}^{-1} k_{k-1} - (h_k + \frac{\lambda^2 \dot{Q}^{-1}}{1 - \lambda^2} M_{k-1}^{-1} k_{k-1})^H h_k \\ \frac{\lambda^2 \dot{Q}^{-1}}{1 - \lambda^2} + M_{k-1}^{-1} \end{cases} \]  

(37)

The integral (35) is then a constant times \( e^C \), which depends on \( h_k \). The term \( C \) must then be written again into a quadratic expression of \( h_k \).

\[ C = (h_k - j)^H R (h_k - j) + B \]  

(38)

\begin{cases} R = \left( \lambda^2 M_{k-1} + (1 - \lambda^2) \dot{Q} \right)^{-1} \\ j = \lambda k_{k-1} \\ B = 0 \end{cases} \]  

(39)

Together with the term outside the integral (35), this is

\[ p(h_k | y_k, I_k) = \alpha_t e^{(h_k - \lambda k_{k-1})^H M_{k-1}^{-1} (h_k - \lambda k_{k-1})} \]  

(40)

\begin{cases} M_k = \lambda^2 \left( M_{k-1} + \frac{1}{\lambda^2} \dot{Q} \right) \\ \lambda k_{k-1} + \frac{1}{\lambda^2} M_{k-1} P_k (h_k' - \lambda k_{k-1}) \end{cases} \]
And the MMSE estimator \( \hat{h}_k \) for the channel at time index \( k \) is the first order moment of a Gaussian distribution centered in \( k_0 \), which is \( \hat{h}_0 = k_0 \). At initial time instant \( t_0 \), if nothing but the channel delay spread \( L \) is known, \( M_0 = Q \) from the maximum entropy principle, as shown in [12], and \( k_0 = 0 \). Therefore, we prove by the above recursion that, under this state of initial knowledge, for all \( k \in \mathbb{N} \), \( p(h_k|I_k) \) is Gaussian with mean \( k_k \) and variance \( M_k \), and \( h_0 = k_0 \). Note that, while regularized inverses of \( Q \) were used along the derivations, the final formulas are properly conditioned with respect to \( Q \).

Note that the solution in (40) coincides with the classical structure of adaptive filters [16] such as the Kalman filter [15], for the problem of dynamic estimation of \( h_k \) from the observed \( y_k \), when \( h_k \) is assumed to evolve in time as an order-1 auto-regressive model. As such, Kalman filters are compliant with our current framework and are developed in section V-D for the sake of comparison. However, when some system parameters are not perfectly known, Kalman filters depart from our general minimal update framework as it will be further detailed.

C. Imperfect system parameter knowledge

In practical applications, contrary to what was stated in Section V-B, the different parameters \( \lambda \), \( \sigma^2 \), and \( L \) especially, are not perfectly known. For simplicity we assume those parameters are constant over the duration of the channel estimation process. Following the maximum entropy principle, these parameters must be assigned an a priori distribution. Let us focus on the time correlation \( \lambda \), which is typically the most difficult parameter to track. In this respect, one has

\[
p(h_k|y_k, I_k) = \int p(h_k|y_k, \lambda, I_k)p(\lambda|y_k, I_k)\,d\lambda \tag{41}
\]

Since \( y_k \) cannot bring alone any cogent information on \( \lambda \),

\[
p(\lambda|y_k, I_k) = p(\lambda|I_k).
\]

The probability \( p(h_k|y_k, \lambda, I_k) \) was computed in Section V-B and is given by the right-hand side of Equation (39), in which \( \alpha \) depends on \( \lambda \) and must therefore be made explicit.

Further computation leads to

\[
p(h_k|y_k, I_k) = \int p(\lambda|I_k)\alpha(\lambda)e^{(h_k-k_0)^T(M_k^{-1})^{-1}(h_k-k_0)}\,d\lambda \tag{42}
\]

with \( M_k(\lambda) \) and \( k_0(\lambda) \) given by Equation (40) for the \( \lambda \) in question and

\[
\alpha(\lambda) = \beta e^{-x(\lambda)\det[X(\lambda)]} \tag{43}
\]

with

\[
X(\lambda) = \left( I + \frac{P_0}{\sigma^2}(\lambda^2M_{k-1}^{(\lambda)} + (1-\lambda^2)Q) \right)^{-1}
\]

\[
x(\lambda) = (\lambda k_{k-1}^{(\lambda)} - h_k^{(\lambda)})^T X(\lambda)^{-1} \left( \lambda k_{k-1}^{(\lambda)} - h_k^{(\lambda)} \right)
\]

\[
\beta = \left( \int p(h_k|y_k, I_k) \, dh_k \right)^{-1}, \text{ independent of } \lambda.
\]

In particular, the conventional MMSE estimate \( \hat{h}_k^{(\text{MMSE})} \) is then the weighted sum

\[
\hat{h}_k^{(\text{MMSE})} = \int \frac{p(\lambda|I_k)e^{-x(\lambda)\det[X(\lambda)]}k_0(\lambda)}{p(\lambda|I_k)e^{-x(\lambda)\det[X(\lambda)]}} \, d\lambda \tag{45}
\]

This integral is however very involved. In practice, it must be broken into a finite sum over a set of potential values for \( \lambda \). Denoting \( S \) this set and \( |S| \) its cardinality, the recursive algorithm that provides the successive estimates \( h_k^{(\lambda)} \), \( k = 1, \ldots, K \), requires that at every step, the values for \( M_k^{(\lambda)} \) and \( k_0^{(\lambda)} \), \( \lambda \in S \) are kept in memory.

Note that the MMSE estimator (45) is no longer linear in \( h_k \) and, as such, does no longer enter the conventional linear Kalman filters. In the next section, we propose simulation results for the proposed minimum update channel estimators, and compare them to maximum entropy channel estimators derived in [7]. We study hereafter classical approaches and show how they differ from or are special cases of our derived techniques.

D. Comparison with classical channel estimation techniques

The channel estimation problem is related to the channel model assumed, mainly determined by the electromagnetic propagation characteristics of the wireless transmission such as transmission bandwidth, carrier frequency, relative speed and spatial configuration of the propagation environment which itself rules the multipath.

These conditions characterize the channel correlation function in a two-dimensional space comprising frequency and time domains. In the general case, each multipath channel component can experience different but related spatial scattering conditions leading to a full bi-dimensional correlation function across these domains.

Nevertheless, classical Clarke and Jakes derivations [18] [19] are based on the assumption that the physical scattering environment is chaotic and therefore the angle of arrival of the electromagnetic wave at the receiver is a uniformly distributed random variable in the angular domain. As a consequence, the time-correlation function is strictly real-valued and governed by the well known expression \( r_v(\Delta t) = J_0(2\pi f_v\Delta t) \) where \( J_v \) is the zero\(^{th}\)-order Bessel function. In addition, the Doppler spectrum is symmetric and interestingly there is a delay-temporal separability property in the general bi-dimensional scattering function.

Under this light, the Wide-Sense Stationary Uncorrelated Scattering (WSSUS) channel model has been proposed [17] and commonly employed for the multipath channels experienced in mobile communications.

This framework might be suboptimal in general. For example, when the mobile is moving in a fixed and known direction, as for example in rural or suburban areas, the WSSUS model would be non applicable. Instead, it can be considered to be separable when the direction of motion averages out because each multipath component is the result of omnidirectional scattering from objects surrounding the mobile, as one would expect in urban and indoor propagation scenarios. Separability is a very important assumption for reducing the complexity of channel estimation, allowing the problem to be separated into two one-dimensional operations.

Hence, for the sake of comparison with the methods proposed in previous paragraphs and which do not make the separability assumption, the channel can be estimated using
a two step approach. First, pilot sub-carriers are used to estimate the whole channel impulse response (CIR) performing frequency-smoothing on each OFDM symbol where pilots are present. Secondly, the smoothed impulse response functions corresponding to a set of OFDM symbols is used in order to improve the channel transfer (CTF) function estimate at the symbol of interest.

Even though TD filtering could be applied remaining in the frequency domain to CTF estimates rather than to the CIR in the time domain because of the linearity relationship between the two, we prefer this option to limit the complexity of the operation.

Thus, for the first step, the Frequency-Domain (FD) optimal MMSE estimator under the assumption of uniform channel power delay profile of known length \( L \) is linear and given by

\[
\hat{\nu}_k^{(FD)} = (F_k^H P_k^H P_k F_L + \sigma^2 I_L)^{-1} F_k^H P_k v_k
\]  

(46)

For the second step, Time-Domain (TD) filtering to exploit time correlation with the channel at previous OFDM symbols containing pilots can be approximated in the form of a finite impulse response filter.

The channel CIR at the \( l \)th tap position and at time instant \( k \) is estimated as

\[
\hat{\nu}_{l,k} = w_l^H \nu_{l,k}^M
\]  

(47)

where we exploit the vector \( \nu_{l,k}^M = [\hat{\nu}_l^{(FD)}, \ldots, \hat{\nu}_{l,k-M+1}^{(FD)}]^T \) of length \( M \) of \( l \)th tap estimates across \( M \) time instants.

Finite length filter approximation seems reasonable as the correlation between consecutive symbols decreases as the terminal speed increases. The fact that the TD correlation is inversely proportional to the terminal speed sets a limit on the possibilities for TD filtering in high-mobility conditions.

The statistical TD filter which is optimal in terms of Mean Square Error (MSE) [16] is the \( M \times 1 \) vector \( w_l \) given by

\[
w_l = (R_{v_k} + \sigma^2 I)^{-1} r_{v_k}
\]  

(48)

where \( R_{v_k} = E[\nu_{l,k}^M(\nu_{l,k}^M)^H] \) is the \( l \)th channel tap \( M \times M \) correlation matrix and \( r_{v_k} = E[\nu_{l,k}^M v_{n,l}] \) the \( M \times 1 \) correlation vector between the \( l \)th tap of the current channel tap realization and \( M \) previous realizations including the current one.

In practical cases, the FIR filter length \( M \) is dimensioned according to a performance-complexity trade-off as a function of the terminal speed.

As an alternative to FIR TD-MMSE channel smoothing coefficient computation, an adaptive estimation approach can be considered which does not require knowledge of second-order statistics of both channel and noise. A feasible solution is the Normalized Least-Mean-Square (NLMS) estimator.

It can be expressed exactly as in Equation (47) but with the \( M \times 1 \) vector of filter coefficients \( w \) updated according to

\[
w_{l,k} = w_{l,k-1} + \mu c_{l,k} e_{l,k}
\]  

(49)

where \( M \) here denotes the NLMS filter order. The \( M \times 1 \) update gain vector is computed according to the well-known NLMS adaption

\[
c_{l,k} = \frac{\mu}{\| \hat{\nu}_{l,k}^M \|^2} \hat{\nu}_{l,k}^M
\]  

(50)

where \( \mu \) is an appropriately-chosen step adaption and

\[
e_{l,k} = \hat{\nu}_{l,k} - \hat{\nu}_{l,k-1}
\]  

(51)

It can be observed that the TD-NLMS estimator requires much lower complexity compared to TD-MMSE as no matrix inversion is required, as well as not requiring any a priori statistical knowledge.

Finally, using both MMSE or NLMS approaches, the CTF channel estimate at \( k \)th symbol can then be retrieved by

\[
h_k = F_L \nu_k
\]  

(52)

When the channel \( h_k \) is modelled in a similar manner as previous paragraph, i.e. the channel evolution across time is expressed by the following state-space model

\[
\begin{align*}
h_k &= \lambda h_{k-1} + \sqrt{1-\lambda^2} w_k \\
h_k' &= h_k + n_k
\end{align*}
\]  

(53)

where \( w_k \sim \mathcal{C}\mathcal{N}(0, \Omega) \) is known as the channel innovation term and \( n_k \sim \mathcal{C}\mathcal{N}(0, \sigma^2 I_N) \) is the additive white Gaussian noise on pilot (and data) symbols. It is to be noted that this 1st order auto-regressive model still complies with the statistical assumption on \( h_k \sim \mathcal{C}\mathcal{N}(0, \Omega) \) made so far.

Under these assumptions, one can easily come up with the expression of a channel estimator according to the classical Kalman form [15]. In fact, letting \( F(L)_{n,n} = e^{-\frac{2\pi i}{L} n} \) with \( 0 \leq n \leq N-1 \) and \( 0 \leq m \leq L-1 \), this can be written as

\[
M_k = F(L) \left( \lambda^2 C_{k-1} + \frac{1}{\lambda^2} I_L \right) F_L^H F_k^H
\]  

\[
\times \left( P_k F_L \left( \lambda^2 C_{k-1} + \frac{1}{\lambda^2} I_L \right) F_L^H P_k^H + \sigma^2 I_N \right)^{-1}
\]  

\[
k_k = \lambda k_{k-1} + M_k P_k (h_k' - \lambda k_{k-1})
\]  

\[
C_k = (I_L - K_k P_k^2 F_L) \left( \lambda^2 C_{k-1} + \frac{1}{\lambda^2} I_L \right)
\]  

(54)

Interestingly, as previously pointed out in Section V-B, Equation (54) is somewhat consistent with (40) derived using the minimal update approach. Nevertheless, both expressions differ in the adaption mechanism which, in the case of the Kalman estimation algorithm, relies on the Kalman-gain \( M_k \) and the error estimate covariance matrix \( C_k \) update. Notice importantly that the estimation process assumes the knowledge of noise statistics \( \sigma^2 \) and the channel length \( L \). In case the latter is not provided, it would be necessary to assume \( L \) as the largest channel length allowed by the OFDM system parameters in use. Therefore in practice, one should dimension it as \( L = \lfloor N/M \rfloor \) [2] in case of imperfect knowledge. In spite of these limitations, the Kalman estimator is often chosen because of the well known robustness against non-stationarity of the signal statistics via the adaptation of the estimate covariance matrix. In order to counter the intrinsic need of parameter knowledge, one could think of using Expectation Maximization in conjunction with Kalman or plain MMSE techniques. Indeed, with an additional complexity cost, any of such channel estimator can be coupled with parameter estimation (speed or channel length) in an iterative fashion. Nevertheless, contrary to the original methods presented based on Maximum Entropy principle and then constructed to be robust with respect to parameter knowledge, they would need the necessary amount of data to converge to construct the
correct a-priori information. Such methods are then well suited only in those cases where the channel is stationary.

Note that other classical adaptive estimators such as (normalized) least mean squares and recursive least squares, that discard most a priori knowledge, perform much less accurately than optimal 2-D MMSE optimal filter [16].

VI. SIMULATION AND RESULTS

In this section, we provide simulation plots to compare, at time \( t \), the minimal channel estimation update method against (i) the one-dimensional MMSE [3], [2], taken at time \( t \), which takes only into account the last past pilot symbols and uses a fixed empirical covariance matrix, (ii) the optimum two-dimensional MMSE provided in [12], (iii) the 1D+1D optimum MMSE, (iii) the 1D+NLMS and (v) the Kalman provided as reference in the section of classical channel estimation techniques, with \( K = 4 \) pilot time indexes. The OFDM DFT size is \( N = 64 \), the channel length \( L = 6 \) is known to the receiver, the vehicular speed is \( v = 120 \) km/h, pilot sequences are transmitted every 0.29 ms (as in 3GPP-LTE [14]), and the induced Jake’s time correlation \( \lambda \) between \( t \) and the past pilot sequence arrival time is known to the receiver. In scenario (ii), all \( K \) past received pilot sequences and time correlations are perfectly known. The channel time correlation model is a \( K \)-order autoregressive model following [13]. A performance comparison is proposed in Figure 2. We notice here that the minimal update algorithm does not show significant performance decay compared to the optimal two-dimensional MMSE estimator, while the one-dimensional MMSE estimator, also relying on the last past pilot sequence, shows large performance impairment. Kalman estimation shows to be comparable in performance only when the channel length parameter is perfectly known but heavily impaired when the maximum channel length assumption is taken instead. The 1D+1D MMSE shows to behave exactly as optimal 2D as well, only when perfect knowledge of parameters is assumed. Interestingly, the NLMS method shows to fail because of the extremely little adaptation lag used in this comparison. Anyway, results not presented here show that NLMS can only be useful if allowed to train over long periods of hundreds of symbols.

In Figure 3, with the same assumptions as previously, we consider the hypothesis where the vehicular speed \( v \) is a priori known to be (with equal probabilities) either 5, 50, 120 km/h. The performance is compared against the optimal 2-D algorithm where \( v \) is known but erroneously estimated (\( v = 5, 50, 120 \) km/h). It is observed that, again, even when \( \lambda \), or equivalently \( v \), is a priori unknown, the Bayesian minimal update framework manages to ideally recover the channel with no performance decay. On the opposite, when \( \lambda \) is erroneously estimated, the performance decay of the optimal estimator might be dramatic.

In Figure 4, we show the Block Error Rate (BLER) performance comparison of a realistic LTE OFDM setup with Turbo Codes and actual signal detection and channel decoding. The BLER plots are obtained for classical OFDM detection performed using the minimal update and Kalman (with imperfect knowledge of channel length parameter) channel estimation. The case where detection is performed using ideal channel knowledge is also presented for the sake of completeness. The minimal update channel estimation provides performances that lies in between the other two cases. Moreover, it shows to offer the same performance as for the optimal 2D estimation although the plot has been omitted for clarity. Hence, the robust minimal update estimation method reveals to be an excellent choice with respect to a method of similar structure and complexity such as Kalman but avoiding the bargain of estimating side information.

VII. CONCLUSION

In this paper, we proposed a novel framework to channel estimation, applied to OFDM-based systems. We successively discussed the fundamental nature of channel estimation, un-
under a Bayesian point of view. This approach allowed us to redefine channel estimation as a technique allowing one to infer the posterior probability distribution $p(h|y, I)$ of a channel $h$ given some input data $y$ and prior information $I$. Assuming the receiver is allowed to store as much past information as desired, we then discussed optimal channel estimators under various levels of prior information at the receiver; the optimality emerges from a systematic usage of the maximum entropy principle. Then we proposed a novel approach to extend the maximum entropy setup when the receiver is oblivious of past received data. In a particular case, the latter was shown to be equivalent to the classical Kalman filter. Simulations suggest that the proposed novel technique is indistinguishable in performance from the optimal infinite time maximum entropy approach.

VIII. ACKNOWLEDGEMENT

This work was partially supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless Communications NEWCOM++.

REFERENCES


Romain Couillet was born in Abbeville, France. He received his Master of Science degree (specialization in Mobile Telecommunications) from the Eurecom Institute, France in 2007. He received his Master of Science degree (specialization in communication systems) from Telecom ParisTech, France in 2007. In September 2007, he joined NXP, founded by Philips, which eventually merged into the ST-Ericsson company. In ST-Ericsson, he works as an algorithm development engineer on the Long Term Evolution Advanced (LTE-A) project. In parallel to his position in ST-Ericsson, he is currently pursuing his Ph.D. degree at Supélec, France. His research interests include wireless communications, multi-user MIMO detection techniques, cognitive radio and random matrix theory. Romain Couillet is the recipient of the Valuetools 2008 best student paper award.

Prof. Méraviana Debbah was born in Madrid, Spain. He entered the Ecole Normale Supérieure de Cachan (France) in 1996 where he received the M.Sc and the Ph.D. degrees respectively in 1999 and 2002. From 1999 to 2002, he worked for Motorola Labs on Wireless Local Area Networks and prospective fourth generation systems (OFDM and MC-CDMA). From 2002 until 2003, he was appointed Senior Researcher at the Vienna Research Center for Telecommunications (fiw., Vienna, Austria) working on MIMO wireless channel modeling issues. From 2003 until 2007, he joined the Mobile Communications department of the Institute Eurecom (Sophia Antipolis, France) as an Assistant Professor. He is presently a Professor at Supélec (Gif-sur-Yvette, France) holder of the Alcatel-Lucent Chair on flexible radio. His research interests are in information theory, signal processing and wireless communications.

Andrea Ancora graduated in 2000 obtaining a “Laurea in Ingegneria Elettronica” from the Politecnico di Torino (Italy) and a degree in Mobile Communications from Institute Eurecom in Sophia Antipolis (France). Andrea is pursuing a PhD program with Eurecom Institute. From May 2000 until now he worked for Texas-Instruments, Philips, NXP and ST-Ericsson in Sophia Antipolis. He is in the Algorithm Development Team for 2G, UMTS/HSDPA and LTE mobile terminal systems. He has been participating to 3GPP standardization. He is inventor of numerous US and European patents.
Adaptive Coded Continuous-Phase Modulations for Frequency-Division Multiuser Systems
Alberto Perotti, Piotr Remlein and Sergio Benedetto

Abstract—In this paper we study adaptive coded modulations for wireless channels. Nonlinearity due to the radio-frequency power amplifier is considered and continuous-phase modulations (CPM) are adopted in order to make the nonlinearity effects negligible. The structure and properties of CPM are reviewed in a semi-tutorial fashion and coded modulation schemes are proposed. Moreover, frequency-division multiplexing (FDM) multiuser systems employing coded CPM are considered. Modulation-coding schemes achieving increasing spectral efficiencies in a multiuser scenario with tight intercarrier frequency spacing are designed and their performance in terms of spectral efficiency and error rate is assessed.

Index Terms—Continuous-phase modulation, coded, multiuser, adaptive, cancellation, iterative, nonlinear.

I. INTRODUCTION

Among the most relevant challenges in the design of wireless transceivers are those concerning the physical layer. It consists of an analog front-end (AFE), including a baseband and radio-frequency (RF) analog processing subsystem and a digital baseband processor. At the transmitter, the AFE performs digital to analog conversion and frequency conversion from baseband RF. The converted signal is then amplified and sent to the antenna. At the receiver side, the AFE consists of the analog subsystem that amplifies the received signal, converts it to baseband and then to digital samples.

The impairments caused by the AFE on the transmitted signal include the distortion caused by the nonlinear characteristic of the transmitter’s RF high power amplifier (HPA), which is known to cause a possibly significant degradation. HPAs are often driven close to or at saturation in order to achieve the highest possible transmitted power. In such case, however, their nonlinear input-output characteristic results in a transmitted signal affected by several kinds of degradations, such as intermodulation distortion, spectral widening, amplitude and phase distortion.

These impairments can be easily explained considering the memoryless nonlinear model [1]–[3] often used to characterize the HPA. According to this model, the complex envelope of the output signal is

\[ x(t) = A \left| r(t) \right| \exp \left\{ j \Phi \left( \left| r(t) \right| \right) \right\} \left| r(t) \right| \]  

where \( r(t) \) (resp. \( x(t) \)) is the complex envelope of the HPA input (resp. output) signal, \( A \left| r(t) \right| \) is the input-output amplitude characteristic (the AM-AM conversion characteristic), and \( \Phi \left( \left| r(t) \right| \right) \) is the input-output phase characteristic (the AM-PM conversion characteristic).

When viewed in the frequency domain, the nonlinear transformation experienced by the input signal corresponds to a possibly multiple convolution of \( R(f) = F[r(t)] \) with itself. Hence, the power spectral density (PSD) of \( x(t) \) is spread over a larger frequency range, thus resulting in a wider spectrum.

When viewed in time domain, the HPA’s nonlinear memoryless transformation applied to the sum of two monochromatic signals with distinct frequencies, \( f_1 \) and \( f_2 \), in the signal bandwidth causes an unwanted amplitude modulation which results in infinite monochromatic components – the intermodulation products – at the PA output, each characterized by a frequency \( f_{ij} = if_1 + jf_2 \), with \( i, j \in \mathbb{Z} \). Some of these components fall into the signal bandwidth, hence causing a kind of self-interference called intermodulation.

A possible approach to overcome these impairments is predistortion (see [4] and references therein). This approach requires an accurate knowledge of the PA characteristic and the implementation of a suitable predistortion circuit, which adds complexity to the system. The key observation that both the AM-AM conversion and AM-PM conversion in (1) depend on the amplitude \( \left| r(t) \right| \) of the input signal leads to the following consideration: an input signal whose complex envelope exhibits constant-amplitude would eliminate the impairments due to AM-AM conversion and result in a fixed phase rotation of the complex envelope, which could be easily recovered by the receiver’s carrier phase recovery circuit without requiring further compensation.

This issue is perhaps the main motivation that led to the adoption of Continuous-Phase Modulations (CPM) [5], a rich class of bandwidth-efficient modulation techniques featuring the constant envelope property. In fact, as shown in Fig. 1, while the complex envelope of a CPM signal has a constant amplitude (see Fig. 1), the complex envelope of a typical QPSK signal exhibits deep amplitude variations.

In the rest of this paper, we will focus on the study of constant-envelope CPM signals and coded CPM systems in a multiuser uplink scenario. The case of many consumer-grade terminals equipped with a low-cost transmitter AFE is interesting, e.g., in the context of uplink satellite systems like the Digital Video Broadcasting (DVB) standard’s Return Channel to Satellite (RCS) [6].

This work was supported by the European Commission through the Network of Excellence in Wireless COMmunications NEWCOM++) (contract n. 216715).
Alberto Perotti and Sergio Benedetto are with Politecnico di Torino, Dipartimento di Elettronica, Corso Duca degli Abruzzi, 24 - I-10129 - Torino (Italy). Piotr Remlein is with Poznan University of Technology, Faculty of Electronics and Telecommunications, Piotrowo 3a; 60-965 Poznan (Poland).

\[ 1 \text{Here, } F \text{ denotes the Fourier transform} \]
CPM has long been considered a complex modulation with poor spectral efficiency. In fact, due to the constant-envelope constraint, the spectral efficiency of simple CPM schemes like the minimum-shift keying (MSK) [5], is low. Recently, however, new results presented in [7]–[9] show that CPM modulations can achieve high spectral efficiencies, at the cost of a possibly higher receiver complexity. In fact, CPM modulators are devices with memory and, as we will see in detail in the following sections, their maximum-likelihood (ML) receiver is exponentially complex in the modulator’s memory size which is, in turn, inversely related to the bandwidth occupied by the modulated signal. As a consequence, only CPM schemes with low complexity, and therefore large bandwidth, have been considered in the past for practical implementations. In the GSM standard, for example, the MSK scheme, although with a modified (Gaussian) frequency pulse, has been adopted.

Multiuser CPM systems in which users share the same spectrum [10], [11] or feature a tight intercarrier frequency spacing [12] have been studied recently. In [10], multiuser CPM systems with zero frequency and phase offset (downlink assumption) over Rayleigh fading channels have been studied. It has been shown that, using an antenna array at the receiver and an optimum multiuser detector, such systems perform very close to systems free of interference. In [11], an asynchronous multiuser CPM system wherein all users share the same carrier frequency has been considered. Optimum and reduced-complexity multiuser detectors based on the Viterbi algorithm have been proposed. In [12], the authors considered a multiuser SCCPM system in which users are allowed to have individual energy levels, carrier frequencies and phases. The authors show that, letting the code word size grow to infinity, the bit error probability vanishes when the channel SNR is sufficiently large.

In this paper, we will show that it is possible to design adaptive coded CPM systems achieving a high spectral efficiency. We will show that even the simplest CPM schemes, like the MSK, achieve a relatively high SE when used in an FDM context with low intercarrier frequency spacing. To obtain an improved SE, interference cancellation is performed at the receiver. Interference cancellation results essential in multiuser coded FDM-CPM adaptive systems in which all users are allowed to choose the best modulation and coding (ModCod) format according to the channel conditions. In fact, the PSD of a CPM signal has infinite support and therefore the amount of interchannel interference (ICI) experienced by each user link depends on the ModCod used in the adjacent links. In such a scenario, if ICI cancellation is not performed, the choice of the optimal ModCod scheme for each link, a task performed at the receiver by means of a suitable adaptivity algorithm, would result in a very complex optimization problem. Using ICI cancellation, each user link becomes virtually immune to the interference caused by its frequency-adjacent links. As a result, ModCod selection can be performed on a link-by-link basis, thus greatly simplifying the adaptivity algorithm.

In our analysis, we will make the following assumptions: an FDM-CPM system with a fixed and tight intercarrier frequency spacing is considered. Moreover, we allow each user to have individual phases and delays, resulting in an uplink assumption. Ideal power control is assumed, therefore the signal power at the receiver is equal for all users. We show that it is possible to obtain significant improvements in SE by using a simple iterative ICI cancellation technique at the receiver. The resulting receiver complexity is slightly increased by the ICI canceller, although it only grows linearly with the number of users.

The paper is organized as follows: in Sec. II, we review the properties of CPM signals and the structure of CPM modulators and receivers. In Sec. III, we introduce in detail the adopted system model. Sec. IV describes the multiuser receiver including the proposed iterative ICI cancellation algorithm for uncoded CPM. In Sec. V, the channel coding scheme is described and the iterative ICI cancellation algorithm is embedded into the iterative decoding process. Sec. VI describes the design of the adaptive coding and modulation schemes. Sec. VII presents the obtained results.

II. STRUCTURE OF CPM SIGNALS AND MODULATORS

A continuous-phase, constant envelope modulated signal can be written as

\[ x(t) = \sqrt{\frac{2E_s}{T}} e^{j\psi(t)} \]  

(2)

where \( E_s \) is the symbol energy, \( T \) is the symbol interval and

\[ \psi(t) = 2\pi h \sum_{n=-\infty}^{\infty} a_n q(t - nT). \]  

(3)

The phase process \( \psi(t) \) depends on the input information symbols \( a_n \in \{ \pm 1, \pm 3, \ldots, \pm (M - 1) \} \), where \( M = 2^m \) is the size of the input alphabet and \( h = Q/P \) is the modulation index (\( Q \) and \( P \) are relatively prime integers). The phase pulse \( q(t) \) is a continuous function whose support extends over \( L \) consecutive symbol intervals\(^3\) and exhibits the following

\(^3\) When \( L = 1 \), the resulting CPM scheme is full-response, while for \( L > 1 \) we have a partial response scheme.
properties
\[ q(t) = \begin{cases} 0 & t \leq 0 \\ \frac{1}{T} & t \geq LT \end{cases} \]

The phase pulse is usually defined as the integral of a frequency pulse \( s(t) \)
\[ q(t) = \int_{-\infty}^{t} s(\tau) d\tau. \]

Different frequency pulse shapes have been proposed [5]. The most commonly used are the rectangular (REC) and the raised-cosine (RC) frequency pulses. A CPM scheme is then defined by specifying its parameters \( M, h, L \) and \( s(t) \).

When a signal like (2) enters a nonlinear device, whose characteristic is modeled as in (1), it undergoes amplitude and phase distortion. However, since its amplitude
\[ |r| = \sqrt{\frac{2E_s}{T}}, \]

is constant, the amplitude distortion results in a gain factor \( A[\sqrt{2E_s/T}] \) and the phase distortion results in a constant phase shift \( \Phi[\sqrt{2E_s/T}] \), which is compensated for by the carrier phase recovery circuit at the receiver.

A nice feature of CPM signals is their bandwidth efficiency: although the PSD of CPM signals has infinite support, it exhibits a large main lobe centered around the carrier frequency with rapidly decreasing tails. The width and height of the main lobe are related to parameter \( L \), also called the correlation length. In fact, a large \( L \) induces a correlation between adjacent symbols in the modulated process. Since in this case the autocorrelation has a larger support, the signal power is more concentrated around the carrier frequency, resulting in a narrower bandwidth. The bandwidth efficiency results increased but, as we will see later, high values for \( L \) result in a higher receiver complexity.

The properties of CPM have been the subject of several studies in the past. Some key advances have been performed after introducing new CPM signal representations, like the Laurent’s decomposition (LD) [13], which is one of the most cited and exploited in the scientific literature. The LD writes a binary CPM as the sum of amplitude-modulated pulses each having distinct durations and shapes. As a result, a continuous-phase modulator can be implemented as a bank of pulse-amplitude modulators, each using a different pulse, whose outputs are added to form the continuous-phase modulated signal. The success of the LD during the early days of CPM is perhaps due to a characteristic that was frequently exploited for the design of reduced-complexity modulators and receivers. In fact, although the number of pulses resulting from the LD may be very large, an accurate approximation can be obtained using only a small subset of them, i.e., those with the largest energy. This result has been exploited in [14], where a low-complexity coherent receiver consisting of a bank of matched filters followed by a Viterbi processor has been proposed. In [15], it has been exploited to design low-complexity receivers for SCCPM systems. The LD has been extended to non binary CPM in [16]. In [17] it has been used to derive capacity bounds on multiple-access channel theory.

Another popular representation has been proposed in [18] and it is known as the Rimoldi decomposition (RD). According to this result, the CPM modulator can be decomposed into the cascade of a time-invariant convolutional encoder operating on rings of integers (the continuous-phase encoder, CPE) and a memoryless modulator (MM) (see Fig. 2). A relevant feature of this representation is the separation of the modulation process into two subprocesses, the first of which accounts for memory across adjacent symbols. In fact, the CPE is a device with memory, while the MM performs a memoryless mapping between symbols at the CPE output and complex waveforms whose support extends over one symbol interval. The CPE is a convolutional encoder operating on integer rings \( \mathbb{Z}_M \) and \( \mathbb{Z}_P \). Its trellis representation has \( N_s = PM^{L-1} \) states and \( N_c = PM^{L} \) edges. At its output, vectors \( c_n \in (\mathbb{Z}_M)^L \times \mathbb{Z}_P \) are generated. Finally, the MM maps these vectors to the complex waveforms of the MM set \( \Sigma_{MM} \).

The maximum-likelihood (ML) receiver for the modulator shown in Fig. 2 consists of a bank of filters matched to the MM signal space (the space spanned by the signals in \( \Sigma_{MM} \)) followed by a trellis processor matched to the CPE trellis. The matched filters provide to the trellis processor a sufficient statistics for the decoding of CPM signal, which is usually implemented using the Viterbi algorithm, for maximum-likelihood (ML) detection or the BCJR algorithm [19] for maximum a-posteriori (MAP) detection. In coded systems, the BCJR algorithm is usually executed within a SISO [20] block that computes the extrinsic information needed to implement an iterative turbo-like [21] receiver.

A comparison of the Laurent and Rimoldi results shows that both representations require a large set of waveforms, whose size grows exponentially with \( L \) but, while the Laurent waveforms have different energies, the Rimoldi waveforms have equal energy \( E_s \). This characteristic made it difficult to use the RD for the design of low-complexity receivers for long time, until the Authors of [22] proposed an orthogonalization technique based on principal components (PC) and applied it to the MM waveforms. In brief, given a signal space with finite dimension like the one spanned by the MM waveforms, PC orthogonalization provides the orthonormal basis whose truncation to fewer dimensions results in minimum error energy. Applying the PC technique to the MM signal space significantly reduces the number of matched filters while the resulting degradation is negligible.

Fig. 3 shows the structure of a CPM receiver with filter bank designed according to the PC orthogonalization procedure. The filter outputs are sampled at instants \( t = nT \) to
obtain a sufficient statistics $y_n \in \mathbb{C}^{N_{PC}}$ needed for ML or MAP decoding. Here, $N_{PC}$ is the number of complex filters resulting from the PC orthogonalization. The log-likelihood ratios of the MM waveforms $s_i \in \Sigma_{MM}$ are then computed as

$$(\lambda_n)_i = \log \frac{P(s_i|y_n)}{P(s_0|y_n)}$$

where $(\lambda_n)_i$ indicates the $i$th element of vector $\lambda_n$. In order to compute the LLRs, the receiver relies on the knowledge of an estimate $\hat{\rho}$ of the symbol energy to noise ratio $E_S/N_0 = \rho$. Clearly, when $\hat{\rho} = \rho$ the receiver provides optimal outputs, while a possible estimation error causes the receiver to operate in the so-called mismatched condition [23] which results in performance degradation.

The MAP detector computes the signal-wise, symbol-wise and bitwise a-posteriori LLRs

$$(\sigma_n)_s = \log \frac{P(s_s|y_n)}{P(s_0|y_n)}, \quad s = 0, \ldots, |\Sigma_{MM}| - 1$$

$$(\mu_n)_m = \log \frac{P(a_m|y_n)}{P(a_0|y_n)}, \quad m = 0, \ldots, M - 1$$

$$(\beta_n)_d = \log \frac{P(b_d=1|y_n)}{P(b_d=0|y_n)}, \quad d = 0, \ldots, \log_2(M) - 1$$

where $y = (y_n)_{n=1}^{N}$ is the sequence of symbols at the sampled output of the PC filter bank and $N$ is the observation length.

Recently, the RD with PC orthogonalization has been exploited for the design of coded CPM systems: a scheme consisting of an outer convolutional encoder (CE) whose coded bits enter an interleaver and then the CPE has been proposed [24]. This way, the outer CE and the CPE form a serial concatenation similar to what is known in the literature as a serially-concatenated convolutional encoder (SCCC) [25]. Iterative decoding can thus be performed, and it yields rather good performance as shown in [24], [26], [27]. Fig. 4 shows a scheme of the SCCC iterative decoder. $\xi^{(i)}$ is the extrinsic information computed at the $i$th iteration by the SISO block that operates on the CPE trellis.

III. SYSTEM MODEL

We consider a system consisting of a set of $2K + 1$ users transmitting their coded CPM signals $x_k(t)$ to a base station or satellite system equipped with a multireceiver system. The channel is affected by additive white Gaussian noise (AWGN) with two-sided power spectral density $G_n(f) = N_0/2$. To adequately model such uplink systems, signals received from the $K$ users are assumed to be asynchronous and not in phase. Moreover, since our goal is the assessment of the achievable spectral efficiency and error rate performance, we assume that power control is ideal, hence the received signals have equal energy.

The received signal model has the following complex baseband representation:

$$y(t) = \sum_{k=-K}^{K} x_k(t-\tau_k)e^{j(2\pi k\Delta f t + \varphi_k)} + n(t)$$

where $\tau_k$ and $\varphi_k$ are, respectively, the delay and phase affecting the $k$th user’s signal. Moreover, $\Delta f$ is the intercarrier frequency spacing and $n(t)$ is the additive white Gaussian noise.

The signal received from the $k$th user is a continuous phase, constant envelope modulated carrier

$$x_k(t) = \sqrt{\frac{2E_s}{T}} e^{j\psi_k(t)}$$

where $E_s$ is the symbol energy, $T$ is the symbol interval and

$$\psi_k(t) = 2\pi h \sum_{n=-\infty}^{\infty} a_n k q(t-nT)$$
The phase process \( \psi_k(t) \) depends on the input information symbols \( a_{n,k} \in \{ \pm 1, \pm 3, \ldots, \pm (M-1) \} \), where \( M = 2^m \) is the size of the input alphabet.

We assume that all users adopt the same CPM scheme, hence the CPM parameters \( h \) and \( q(t) \) do not depend on \( k \). Although this assumption does not match a real scenario, it becomes reasonable in the absence of ICI resulting from perfect cancellation.

In the following, we will use the shorthand notation
\[
x_k(t) = M(a_k)
\]
where \( a_k \) denotes the symbol vector of user \( k \) and \( M(\cdot) \) denotes the CPM modulation function.

In [28], it has been observed that CPM schemes with rectangular (REC) frequency pulses exhibit good performance when used in FDM-CPM systems. Therefore, in this paper we will consider REC frequency pulses.

IV. THE MULTIUSER RECEIVER

In the model defined in Sec. III, the capacity of the \( k \)th user link is determined by its signal energy to noise ratio \( E_s/N_0 \) and by the ICI. The capacity reduction due to ICI may be significant when, attempting to achieve higher spectral efficiency (SE), the intercarrier frequency spacing is set to low values. In [28], it has been shown that, in such case, the optimal frequency spacing that maximizes the SE when single-user receivers are employed depends on the parameters of the chosen CPM scheme.

The multiuser receiver we propose improves the system SE through iterative ICI cancellation. It consists of a bank of single-user receivers that iterate with an interference-cancellation block which performs simple linear operations on the received signal. The complexity of this receiver grows only linearly with the number of users. Moreover, as shown later, the proposed receiver exhibits significantly improved SE both for uncoded and coded CPM systems.

A scheme of the proposed receiver is shown in Fig. 5. Its main building block is a single-user MAP detector and remodulator (detailed in Fig. 6). Such detector computes the \( k \)th user’s \( a \)-posteriori signal LLRs \( \sigma_{a_k} \) as in (4). To compute the \( a \)-posteriori LLRs, the detector relies on the assumption that the signal being detected is corrupted by AWGN and on the knowledge \( \hat{\rho} \) of the signal-to-noise ratio \( \rho = E_s/N_0 \).

Clearly, \( \hat{\rho} = \rho \) yields the optimal MAP detector for a single-user communication (i.e., no ICI) through the AWGN channel, and bank of \( 2K + 1 \) single-user detectors yields optimal performance for \( \Delta_f \to \infty \) since ICI becomes negligible in such case. In the latter case, however, the resulting SE becomes very low due to the large intercarrier frequency spacing. Since we are interested in spectrally efficient systems, we will consider the case of tight normalized intercarrier frequency spacing, i.e. \( \Delta_f T \lesssim 1 \). As shown in [28], in such case the strong ICI significantly reduces the information rate with respect to a single-user system with no ICI. However, the low \( \Delta_f T \) values compensate for the lower information rate, resulting in a possibly large SE.

The \( k \)th single-user detector and remodulator is fed with the following input signal at the \( n \)th iteration
\[
y_k^{(n)}(t) = y_k^{(n)}(t + \tau_k)e^{-j2\pi k\Delta_f(t+\tau_k)+\varphi_k}
\]
where \( y_k^{(n)}(t) \) is the received signal after interference cancellation, defined as
\[
y_k^{(n)}(t) = y(t) - \sum_{j=-J}^{J} s_k^{(n)}(t)
\]
where \( J \) is a positive integer parameter related to the number of adjacent interfering signals being canceled. The signal \( s_k^{(n)}(t) \) is
\[
s_k^{(n)}(t) = m_k^{(n)}(t - \pi)e^{j2\pi k\Delta_f(t-\tau_k)+\varphi_k}
\]
PEROTTI et al.: ADAPTIVE CODED CONTINUOUS-PHASE MODULATIONS FOR FREQUENCY-DIVISION MULTIUSER SYSTEMS

(a) Information rates

(b) Spectral efficiencies

Fig. 7. Information rates of MSK systems (a) and spectral efficiencies achievable by coded MSK systems (b) with $\Delta f T = 0.5$.

V. CODED CPM MULTIUSER SYSTEMS

Following the SCCPM approach [24], we connect an outer CE to the CPM modulator through an interleaver. A 4-state, rate 1/2 systematic recursive CE with generators $(7, 5)_8$ is adopted (see Fig. 10). Moreover, in order to achieve lower coding rates, the rate 1/2 CE is extended adding a further feed-forward connection to obtain a rate 1/3 encoder. In this case, the generators are $(7, 5, 3)_8$.

The CE output is punctured using a rate-matching algorithm to achieve variable coding rates: the puncturing algorithm selects all systematic bits and some coded bits according to a regular pattern in order to achieve the desired overall rate.

The interleaver that connects the outer encoder to the CPM modulator is a spread (S-random) interleaver [30]. The spread parameter is set according to the code word length.

The adopted structure of channel decoder results in two iteration loops (see Fig. 8): an inner loop formed by the ICI canceler, the CPE SISO decoder, the receiver front-end and the remodulator; and an outer loop formed by the CPE SISO decoder, the CE SISO decoder, the interleaver (II) and the deinterleaver (II$^{-1}$).

The decoder starts decoding a received code word executing $N_{IC}$ inner iterations. Then, it executes $N_P$ outer iterations. Each outer iteration is followed by an inner iteration. This way, ICI cancellation is performed as part of the decoding iterations and it results in an improved ICI cancellation. In fact, as shown in Fig. 9, perfect cancellation and hence an

where $m_i^{(i)}(t)$ is the re-modulated signal for $l$th user at the $i$th iteration.

Several techniques can be applied to compute the re-modulated signal. In [29], a soft technique has been proposed. It consists in the following steps: starting from the a-posteriori signal LLRs $\sigma_{nk}$ of (4), it is possible to compute the a-posteriori probability distribution $p_{nk}$ of the MM signal set $\Sigma_{MM}$ for user $k$ and symbol interval $n$ as

$$\begin{align*}
(p_{nk})_r &= P\{s_{nk} = s_r | y_k\} = e^{\sigma_{nk}(r)} P\{s_{nk} = s_0 | y_k\} \\
(p_{nk})_0 &= P\{s_{nk} = s_0 | y_k\} = \left[\sum_{r=0}^{\Sigma_{MM} - 1} e^{\sigma_{nk}(r)}\right]^{-1}
\end{align*}$$

The re-modulated signal is then computed as the average signal

$$m_i^{(i)}(t) = \sum_{n=\infty}^{\infty} \sum_{r=0}^{\Sigma - 1} (p_{nk})_r s_r(t - nT) \quad (13)$$

where $i$ denotes the iteration number.

Results show that this ICI cancellation technique improves the SE of uncoded CPM systems. As an example, we show in Fig.7 the information rates and spectral efficiencies achieved by an MSK system with normalized intercarrier frequency spacing $\Delta f T = 0.5$. We note that ICI cancellation significantly improves the SE.
almost single-carrier performance is achieved in some cases. The $E_b/N_0$ gain due to ICI cancellation is about 1.2 dB at BER $= 10^{-3}$ and the curve is as close as 0.2 dB to the single-carrier BER curve.

VI. CHOICE OF THE MODULATION AND CODING SCHEMES

The choice of the modulation and coding (ModCod) schemes for an adaptive communication system is driven by the target SE requirements. Assuming a normalized intercarrier frequency spacing $\Delta_f/T \leq 1$, then the size $M$ of the CPM input alphabet determines the maximum achievable $SE = \log_2 M$ of the uncoded system. A quaternary scheme ($M = 4$) results therefore in a maximum achievable SE larger than 2 bits/s/Hz. Moreover, the channel encoder reduces the SE proportionally to its rate. As a result the asymptotic ($E_b/N_0 \to \infty$) SE is

$$SE_{\infty} \triangleq \lim_{E_b/N_0 \to \infty} SE = \frac{R_C \log_2 M}{\Delta_f T}$$  (14)

where $R_C = k_0/n_0$ is the rate of the punctured convolutional code.

The 4-state recursive CE of Fig. 10 features a systematic output ($c_0$) and two coded outputs ($c_1$ and $c_2$), therefore its rate is 1/3. Its output connection $c_2$ is punctured when coding rates larger than 1/2 are needed.

At the output of the CE, puncturing is applied to the coded bits. The rate of puncturing is defined as $R_p = N_p/D_p$.

Systematic bits are never eliminated, while coded bits are selectively eliminated according to a periodic pattern to achieve the desired rate.

As for the modulation scheme, full-response systems are preferred. Indeed, the use of partial-response systems is motivated by the improved spectral compactness needed in systems with tight bandwidth constraints when ICI is not canceled. Moreover, a large $L$ results in an increased receiver complexity. In our case, ICI is reduced through an appropriate cancellation algorithm. Moreover, a low CPM complexity compensates for the increased complexity due to ICI cancellation.

One binary and one quaternary CPM modulation scheme have been chosen. Their parameters are shown in Tab. I. The binary scheme is aimed at increasing the robustness of the ModCod set at low $E_b/N_0$, while the quaternary scheme provides the high number of signals needed to achieve a high SE when the SNR is larger. The binary CPM is a simple MSK scheme, while for the quaternary case a full response scheme with appropriate modulation index has been chosen.

As for the coding schemes, rate values from 3/8 to 5/6 have been considered. This range has been determined in order to achieve $SE_{\infty}$ from 0.5 bits/s/Hz to 2.5 bits/s/Hz with an intercarrier frequency spacing $\Delta_f T = 2/3$. The parameters of each scheme are shown in Tab. II. Tab. III shows the achieved $SE_{\infty}$ values with the chosen modulation and coding schemes.

VII. RESULTS

The main results of this work are shown in Fig. 11 and Fig. 12. Bit error rates (BER) and frame error rates (FER) of the selected ModCod schemes are provided. A range of $E_b/N_0$ from 2 dB to 18 dB is spanned at BER $= 10^{-4}$ with a spectral efficiency spanning the range from 0.5 bits/s/Hz to 2.5 bits/s/Hz. We observe that the curves corresponding to lower spectral efficiencies - those that use M1 as the modulation - exhibit a steep waterfall behavior, while those corresponding to higher spectral efficiencies, and hence modulation M2, exhibit...
Fig. 11. Error rates of the selected coded modulation modes. Dashed lines refer to frame error rates (FER) and solid lines refer to bit error rates (BER).

Table IV

<table>
<thead>
<tr>
<th>C</th>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>(1, 1, 10)</td>
<td>(5, 4, 10)</td>
</tr>
<tr>
<td>C2</td>
<td>(1, 1, 10)</td>
<td>(8, 4, 10)</td>
</tr>
<tr>
<td>C3</td>
<td>(1, 1, 10)</td>
<td>(8, 4, 10)</td>
</tr>
</tbody>
</table>

a slower decrease with $E_b/N_0$. This degradation is mainly due to the much stronger ICI experienced when M2 is used. In such case, using a partial-response system would perhaps improve the performance, although resulting in an increased complexity. Another solution would be the increase of $\Delta f/T$, but this would reduce the SE.

Table IV shows the parameter values used in the simulations. We compare the obtained results with those presented in [31], where SCCPM systems with SE ranging from 0.75 bits/s/Hz to 2.25 bits/s/Hz are designed. The outer codes used therein are convolutional codes with number of states from 4 to 64. Moreover, extended BCH codes and extended Hamming codes are also used as outer codes. In this case, a suboptimal Chase-Pyndiah decoding algorithm [32] is employed at the receiver. Moreover, the CPM schemes used therein are quaternary partial-response schemes with higher complexity.

The information word length used in [31] is 128 bytes = 1024 bits, a value very close to the length used in this paper, which is 1000 bits. Moreover, in [31] a single-user system is considered and the decoder performs 30 iterations, while our decoder performs only 10 iterations.

The best results reported in [31] show that, for a FER = $10^{-3}$, SE = 0.75 bits/s/Hz is achieved at $E_b/N_0$ close to 2 dB and SE = 2.25 bits/s/Hz is achieved at $E_b/N_0$ close to 10.5 dB. Our scheme achieves SE = 0.75 bits/s/Hz at $E_b/N_0 = 2.5$ dB and SE = 2 bits/s/Hz at $E_b/N_0 = 10.5$ dB with FER = $10^{-2}$.

Although our receiver operates in presence of strong ICI, its performance loss is only slight. Moreover, its complexity is lower since CPM schemes are simpler and the number of iterations is significantly lower.

VIII. Conclusion

We have shown that it is possible to design adaptive coded modulation systems employing CPM to mitigate the nonlinearity degradations of wireless transceivers. A high spectral efficiency can be achieved by performing a suitable design of the coding and modulation modes.

The proposed ModCod set spans a range of spectral efficiencies from 0.5 bits/s/Hz to 2.5 bits/s/Hz over a range of $E_b/N_0$ from 2 dB to 18 dB at BER = $10^{-4}$. The adaptive coded modulation system includes an ICI cancellation algorithm, which improves the SE while resulting in a simplified adaptivity algorithm.

Acknowledgments

This work has been supported by the European Commission through the Network of Excellence in Wireless Communications NEWCOM++ (contract n. 216715).

References


Alberto Perotti (S’01-M’03) received the Laurea in Ingegneria delle Telecomunicazioni in 1999 and the Ph.D. degree in Electronics and Communications Engineering in 2003, both from Politecnico di Torino, Turin, Italy. In 2002, he spent seven months as a visiting scholar at the University of California, Los Angeles, working on the design of serially concatenated convolutional codes. Since 2003, he has been working as an associate researcher in the Electronics Department of Politecnico di Torino. His research interests are in the area of wireless communications. In particular, he is interested in channel coding and modulation techniques for wireless systems.

Pierot Remleir received the M.Sc. and the Ph.D. degrees from Poznan University of Technology (PUT), Poznan, Poland in 1991 and 2002 respectively. He has been employed at PUT since 1992, currently as an Assistant Professor at the Chair of Wireless Communications. His scientific interests cover wireless networks, communication theory, error control coding, cryptography, digital modulation, trellis coded modulation, continuous phase modulation, mobile communications, digital circuits design. Dr. Remleir is the author of more than 60 papers, which have been presented at national and international conferences, and published in the communication journals. He also acts as a reviewer for international and national conference and journal papers. He is the member of The Institute of Electrical and Electronics Engineers (IEEE) Communications Society and IEEE Information Theory Society.

Sergio Benedetto (M’76-SM’90-F’97) is a Full Professor of Digital Communications at Politecnico di Torino, Italy since 1981. He has been a Visiting Professor at University of California, Los Angeles (UCLA), at University of Canterbury, New Zealand, and is an Adjunct Professor at Ecole Nationale Superieure de Telecommunications in Paris. He has co-authored two books on probability and signal theory (in Italian), the book “Digital Transmission Theory” (Prentice-Hall, 1987), “Optical Fiber Communications” (Artech House, 1996), and “Principles of Digital Communications with Wireless Applications” (Plenum-Kluwer, 1999), and over 250 papers in leading journals and conferences. He has taught several continuing education courses on the subject of channel coding for the UCLA Extension Program and for the CEI organisation. He has been Chairman of the Communications Theory Symposium of ICC 2001 and CTT2004, and has organized numerous sessions in major conferences worldwide. Sergio Benedetto has been for five years the Area Editor for the IEEE Transactions on Communications for Modulation and Signal Design, a Distinguished Lecturer of the IEEE Communications Society, and the Chairman of the Communication Theory Committee of IEEE and a Fellow of the IEEE. He is Vice President for Publications of the IEEE Communications Society. S. Benedetto is a Fellow of the IEEE, a member of the Science Academy of Torino, and has received the “Premio Siemens per le Tele comunicazioni” at the International Conference on Communications of Genova in 1973, the “Premio Bianchi” of AEI in 1974, the “Premio Bonavera dell’Accademia delle Scienze di Torino” in 1976, the “Gold medal Award for the Best Paper” in 1991, the “Italgas International Prize for Research and Technological Innovation” in 1998, the Cristoforo Colombo International Award for Communications in 2006, and the IEEE Communications Society Edwin Howard Armstrong in 2008.
Iterative Joint Data Detection and Channel Estimation for DS-CDMA Systems in the Presence of Time-Varying Channels

Erdal Panayirci, Hakan Doğan, and Hakan A. Çirpan

Abstract—In this paper, we present an efficient iterative receiver structure of computationally tractable complexity for joint multiuser detection and multichannel estimation (JDE) of direct-sequence code-division multiple-access (DS-CDMA) systems operating in the presence of time-varying flat fading channel. The time-varying channel is assumed to be modeled according to a piece-wise constant channel model at the receiver. The scheme results from an application of the expectation-maximization (EM) algorithm. The resulting EM-JDE receiver updates the data bit sequences in parallel while the channel parameters are also updated in parallel. The EM algorithm provides a set of free parameters, called weight coefficients, which can be selected to optimize its performance. An optimality criterion is defined and analytical expressions for the corresponding optimized weight coefficients are given. Monte-Carlo simulations of a synchronous scenario show that the proposed JDE receiver have excellent multiuser efficiency and are robust against errors in the estimation of the channel parameters. Moreover, very short training sequences are required for the JDE schemes to converge.

I. INTRODUCTION

MULTIUSER detection is known to drastically increase the bandwidth efficiency of code-division multiple-access (CDMA) systems compared to conventional detection using RAKE receivers [1]. However, the complexity of the optimum multiuser receiver/detection processing grows exponentially with the number of users and the number of multi-paths, which prevents any possibility of implementation [2]. Thus, suboptimum feasible techniques for multiuser detection have been proposed which still approach the performance of the optimum receiver. Worth mentioning among them are linear multiuser detection [3] and iterative cancelation of multi-access interference (MAI) in the received signal before making a data decision. An overview of suboptimum multiuser detection techniques can be found in [4] and [5]. The expectation-maximization (EM) algorithm [6], [7] is an iterative method which enables approximating the maximum-likelihood (ML) estimate when a direct calculation of this estimate is computationally prohibitive.

This research has been conducted within the Network of Excellence in Wireless Communications (NEWCOM) ++ inspired funded through the 7th Framework Programme.

Ererd Panayirci is with the Department of Electronics Engineering, Kadir Has University, Cihalı 34083, Istanbul, Turkey, e-mail: eepanay@khas.edu.tr

Hakan Doğan and Hakan A. Çirpan are with the Department of Electrical and Electronics Engineering, Istanbul University, Avcilar 34850, Istanbul, Turkey, e-mail: {hdogan, hcirpan}@istanbul.edu.tr

Georghiades and Han proposed an EM-based receiver which performs joint data detection and estimation for the time-variant flat Rayleigh fading channel[8]. Feder and Weinstein apply the EM algorithm to the problem of estimating parameters of superimposed signals [9]. Following this approach, Borran and Nasiri-Kenari develop in [10] a computationally feasible multiuser detector for the AWGN channel. Finally, turbo EM algorithm is used in [11] for maximum likelihood sequence detection and estimation(MLSDE). Kocian and Fluery [12] considered joint data detection and channel estimation of MC-CDMA systems in the presence of flat fading channels. Panayirci et al. extended their results to the uplink Multi-carrier CDMA systems with frequency selective channels[13].

This paper is mainly an extension of the work of Kocian and Fluery to the time-varying channels. The basic assumption in their work was that the channel remains constant over whole the observation frame, which is not actually suitable for the cases requiring high mobility. We extended their results to the problem of joint multiuser data detection and channel estimation (JDE) of synchronous direct-sequence (DS)-CDMA signals in the time-varying flat Rayleigh fading channel. The time-varying channel is assumed to be modeled according to a piece-wise constant channel at the receiver. In this way, we obtain iterative methods of tractable complexity which smartly combine the two processes of data detection and channel estimation. The synchronous assumption and the flat-fading channel model provide a simple framework for studying JDE in the uplink of a DS-CDMA system. The JDE algorithms presented in this paper can also be extended to the asynchronous case and to frequency- selective channels.

The paper is organized as follows. The signal model of DS-CDMA system and the time-varying channel model is outlined in Section II, followed by a concise description of the EM technique is given and is applied to derive the so-called EM-JDE receiver in Section III. This scheme encompasses weight coefficients which can be selected to optimize its performance. An optimality criterion is defined and the corresponding optimum weight coefficients are derived. Finally, in Section sec:4, the performance of the JDE algorithms is assessed and compared to that of the minimum mean-square error (MMSE) [3] by means of Monte-Carlo simulations.

Notation: Vectors (matrices) are denoted by boldface lower (upper) case letters; all vectors are column vectors; (.)∗, (.)T and (.)† denote the conjugate, transpose and conjugate transpose, respectively; ∥.∥ denotes the Frobenius norm; IL
denotes the $L \times L$ identity matrix; $\text{diag} \{ \cdot \}$ denotes a diagonal matrix; $\Re \{ \cdot \}$ denotes the real part of $\{ \cdot \}$.

II. SIGNAL MODEL

We consider a synchronous CDMA system with $K$ active users. Let us assume that the channel is time-varying for all users. Therefore, each user experiences different time-varying channel. The baseband representation of the received signal, therefore, can be expressed as

$$ z(m) = RB(m)a(m) + n(m), \quad m = 0, 1, \cdots, M - 1 \quad (1) $$

where $z(m) = [z_1(m), z_2(m), \cdots, z_K(m)]^T$ denote the complex vector whose components represent $K$ received signals of the $m$-th user. The $K \times K$ diagonal matrix $B(m)$ is given by

$$ B(m) = \text{diag} \{ b_1(m), b_2(m), \cdots, b_K(m) \} $$

and $b_k \in \{-1, +1\}$, denoting the binary symbols transmitted by $k$-th user during the $m$-th signaling interval. The vector $a = [a_1(0), \cdots, a_{K}(M-1), a_{K}(0), \cdots, a_{K}(M-1)]^T$ represents the time-varying frequency-nonselective channel where the noise vector $w(m)$ is such that

$$ \text{Cov}(w(m), w(n)) = \sum_{k=1}^{K} \rho_{kk} \delta(m-n) $$

where $\rho_{kk}$ is the complex Gaussian vector.

Now let us assume that the signature waveforms are selected in such a way that the correlation function

$$ r_k(q - q') = \sigma_k^2 J_0(2\pi f_d (q - q')), \quad \text{for} \quad q, q' = 1, 2, \ldots, \Xi. \quad (4) $$

with $r_k$ being the crosscorrelations between the signature waveforms of users $j$ and $k$.

Fig. 1.

While the wireless channel is time-varying frequency-nonselective, the channel matrix can be modeled as a piece-wise constant channel. That is, the channel is assumed to be constant during $L$ consecutive symbols. $L$ is the parameter which may depend on the velocity or the doppler frequency. Thus, the number of channel coefficients to be estimated in an observation frame of $M$ symbols ($M >> L$) is reduced and the quality of the channel estimation algorithm is improved. Following the piece-wise constant channel model, Eq.(1) can be replaced by

$$ z(m) = RB(m)a(\frac{m}{L} + 1) + n(m), \quad m = 0, 1, \cdots, M - 1 \quad (2) $$

where $\lfloor x \rfloor$ denotes the integer which is less than or equal to $x$. Let $q = \lfloor \frac{m}{L} \rfloor + 1$ then $a(q) \triangleq [a_1(q), a_2(q), \cdots, a_K(q)]^T$, $q = 1, 2, \cdots, \Xi$, $a_k \triangleq [a_k(1), a_k(2), \cdots, a_k(\Xi)]^T$ is the reduced vector of the channel coefficients and the integer $\Xi$ is such that $L\Xi = M$. Finally, $n(m)$ in (2) is a $K \times 1$ dimensional zero-mean Gaussian random vector with covariance matrix $N_0 R$.

In general, the time variation of the channel coefficients can be modeled using an autoregressive (AR) model of order $n$. For first-order case, the channel coefficients of each user can be expressed as

$$ a_k(q) = \gamma_k a_k(q - 1) + \epsilon_k(q), \quad k = 1, 2, \cdots, K; \quad q = 1, 2, \cdots, \Xi \quad (3) $$

where $\gamma_k$ is the time correlation coefficients and $\epsilon_k(q)$ is the additive white Gaussian noise (AWGN) with zero mean and variance $\sigma_k^2$. The parameter $\gamma_k$ can be determined based on Jakes’ model [16], [18].

$$ r_k(q - q') = \sigma_k^2 J_0(2\pi f_d (q - q')), \quad \text{for} \quad q, q' = 1, 2, \ldots, \Xi. \quad (4) $$

The complex Gaussian vector $w(m) = (F^T)^{-1}n(m)$ is white, i.e., has covariance matrix $N_0 R$ because of Cholesky factorization. We will adopt employing this description of observation model for the subsequent analysis, since the property of the noise vector $w(m)$ to be white.
The received vector (5) may be expressed in terms of the user's components as follows.

\[ y(m) = \sum_{k=1}^{K} f_k b_k(m) a_k \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) + w(m), \quad m = 0, 1, \ldots, M - 1 \]

(6)

where \( f_k \) denotes the \( k \)th column of \( F \) and \( b_k(m) \) is data sent by the user \( k \) within the \( m \)th signaling interval. Suppose a frame of \( M = L \Xi \) symbols are transmitted. We stack \( y(m) \) as \( y = \left[ y_T^T, y_{T-1}^T, \ldots, y_{1}^T \right]^T \), \( y_i = \left[ y^T(i), y^T(i + L), \ldots, y^T(i + (\Xi - 1)L) \right]^T \), \( i = 0, 1, \ldots, L - 1 \).

Then the received signal model can be expressed in more succinct form

\[ y = \Psi a + w \]

(7)

where \( a = \left[ a_1^T, a_2^T, \ldots, a_K^T \right]^T \), \( a_k = [a_k(1), a_k(2), \ldots, a_k(\Xi)]^T \), \( \Psi = [\Psi_{i,j}] \) is a \( L \times K \) block-matrix whose \( \Xi \times \Xi \)-dimensional block matrices are defined as

\[ \Psi_{i,j} = \text{diag} \{ b_j(i) f_j, b_j(L + i) f_j, \ldots, b_j((\Xi - 1)L + i) f_j \} \]

and \( w \) is likewise defined as vector \( y \).

III. JOINT DATA DETECTION AND CHANNEL ESTIMATION WITH EM ALGORITHMS (EM-JDE)

Let \( \mathbf{b} \) denote a possibly vector-valued parameter to be estimated from some possibly vector-valued observation \( y \) with probability density \( p(y|\mathbf{b}) \). The EM algorithm provides an iterative scheme to approach the ML estimate \( \hat{\mathbf{b}} = \arg \max_{\mathbf{b}} p(y|\mathbf{b}) \) in cases where a direct computation of \( \hat{\mathbf{b}} \) is prohibitive. The derivation of the EM algorithm relies on the concept of a hypothetical, so-called complete unobservable data \( \chi \) which, if it could be observed, would ease the estimation of \( \mathbf{b} \). The observed random variable \( y \) which is referred to as the incomplete data within the EM framework, is related to \( \chi \) by a mapping \( \chi \mapsto y(\chi) \).

The suitable approach for applying the EM algorithm for the problem at hand is to decompose the received vector in (7) into the sum [9]

\[ y(m) = \sum_{k=1}^{K} x_k(m), \quad m = 0, 1, \ldots, M - 1 \]

(8)

where

\[ x_k(m) = f_k b_k(m) a_k \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) + w_k(m). \]

(9)

\( x_k(m) \) represents the received signal component transmitted by the \( k \)th user through the channel with its channel parameter \( a_k(q), q = \left\lfloor \frac{m}{L} \right\rfloor + 1 \). The Gaussian noise vector, \( w_k(m) \) in (9) represents the portion of \( w(m) \) in the decomposition defined by \( \sum_{k=1}^{K} w_k(m) = w(m) \), whose variance is \( N_0 \beta_k \). The coefficients \( \beta_k \) determine the part of the noise power of \( w(m) \) assigned to \( x_k(m) \), satisfying \( \sum_{k=1}^{K} \beta_k = 1 \), \( 0 \leq \beta_k \leq 1 \).

The problem now is to estimate the transmitted symbols \( \mathbf{b} = \{ b_k(m) \}_{k=1,m=0}^{K,M-1} \) and the complex channel responses \( a_k = [a_k(1), a_k(2), \ldots, a_k(\Xi)]^T \) for each user based on observed data \( y \). In the EM algorithm, we view the observed data \( y \) as the incomplete data, and define the complete data as \( \chi = \{ x_1, a_1 \}, \{ x_2, a_2 \}, \ldots, \{ x_K, a_K \} \) where \( x_k = [x_k(0), \ldots, x_k(M - 1)]^T \) with \( M = L \Xi \) and for \( k = 1, 2, \ldots, K \). Given the complete data set, the loglikelihood function of the parameter vector to be estimated \( \mathbf{b} \) can be expressed as

\[ \log p(\chi|\mathbf{b}) = \sum_{k=1}^{K} \left( \log p(x_k|b_k, a_k) + \log p(a_k|b_k) \right) \]

(10)

where, \( x_k = [x_k^T(0), x_k^T(1), \ldots, x_k^T(M - 1)]^T \) and \( b_k = [b_k(0), b_k(1), \ldots, b_k(M - 1)]^T \). We neglect the \( \log p(a_k|b_k) \) term in (10) since the data sequence \( b_k \) and \( a_k \) are independent of each other.

**Expectation Step (E-Step):** The first step to implement the EM algorithm, called the Expectation Step (E-Step), is to compute the average log-likelihood function. The conditional expectation is taken over \( \chi \) given the observation \( y \) and that \( \mathbf{b} \) equals its estimate calculated at \( t \)th iteration.

\[ Q \left( b_i|b^{(i)} \right) = E \left\{ \log p(\chi|b,y, b^{(i)}) \right\} \]

(11)

Taking into account the special form of \( \log p(\chi|b) \) in (10), Eq. (9) can be decomposed as

\[ Q \left( b_i|b^{(i)} \right) = \sum_{k=1}^{K} Q_k(b_k|b^{(i)}) \]

(12)

where

\[ Q_k(b_k|b^{(i)}) = E \left\{ \log p(x_k|b_k, a_k)|y, b^{(i)} \right\} \]

(13)

Note that (13) follows from (10).

Neglecting the terms independent of \( b_k \), from (10), \( \log p(x_k|b_k, a_k) \) can be calculated as

\[ \log p(x_k|b_k, a_k) \sim \sum_{m=0}^{M-1} \Re \{ f_k^T b_k(m) a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) x_k(m) \}. \]

(14)

Inserting (14) in (13), we have for \( Q_k(b_k|b^{(i)}) \)

\[ Q_k(b_k|b^{(i)}) = \sum_{m=0}^{M-1} \Re \{ f_k^T b_k(m) a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) x_k(m) \} \]

(15)

where, adopting the notation used in [12],

\[ (a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) x_k(m))^{(i)} \triangleq E \left\{ a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) x_k(m) | y, b^{(i)} \right\} \]

(16)

The above expectation can be calculated by applying the conditional expectation rule as

\[ a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) x_k(m) = E[a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) x_k(m) | y, b^{(0)}, a]y, b^{(0)} \]

From (9) it follows that the conditional distribution of \( x_k(m) \) given \( y \), \( a \) and \( b = b^{(i)} \) is Gaussian with mean

\[ E[x_k(m) | y, b^{(i)}, a] = f_k^{(0)}(m) a_k^* \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) \]

(17)

\[ + \beta_k \left( y(m) - \sum_{j=1}^{K} b_j^{(i)}(m) a_j \left( \left\lfloor \frac{m}{L} \right\rfloor + 1 \right) \right) \]
where $b^{(i)}_k(m) \triangleq E(b_k(m)|y, b^{(i)}, a)$. Inserting (17) in (18), rewrite (13) can be written as

$$a_k^+(\lceil \frac{m}{L} \rceil + 1) x_k(m) = \mathbf{f}_k b_0^j(m) E\{ |a_k^+(\lceil \frac{m}{L} \rceil + 1) y, b^{(i)} \}$$

$$+ \beta_k E\{ a_k^+(\lceil \frac{m}{L} \rceil + 1) y, b^{(i)} \} y(m)$$

$$- \sum_{j=1}^{K} f_j b_j^j(m) E\{ a_j(\lceil \frac{m}{L} \rceil + 1) |y, b^{(i)} \}. \tag{18}$$

The prior pdf of $a = [a_1^T, a_2^T, \ldots, a_K^T]^T$ is Gaussian and can be expressed as

$$p(a) \sim \exp \left[ -\frac{1}{2} (a - \mu)^T \Sigma^{-1} (a - \mu) \right] \tag{19}$$

where mean and covariance are $m = [m_1^T, m_2^T, \ldots, m_K^T]^T$ and $\Sigma^{-1} = \text{diag} \{ C_1^{-1}, C_2^{-1}, \ldots, C_K^{-1} \}$ respectively.

Furthermore, since $w \sim \mathcal{N}(0, N_0 I)$ form (7) the conditional pdf of $a$ given $y$ and $b^{(i)}$ can be written as follows

$$p(a|y, b^{(i)}) \sim p(y|a, b^{(i)}) p(a)$$

$$\sim \exp \left[ -\frac{1}{N_0} (y - \Psi^{(i)} a)^T (y - \Psi^{(i)} a) - (a - \mu)^T \Sigma^{-1} (a - \mu) \right].$$

After some algebra it can be shown that

$$p(a|y, b^{(i)}) \sim \mathcal{N}(\mu^{(i)}, \Sigma^{(i)}) \tag{20}$$

where

$$\mu^{(i)} = \Sigma^{(i)} \left[ \frac{1}{1} \Sigma^{-1} m + \frac{1}{1} \Psi^{(i)} y \right]$$

$$\Sigma^{(i)} = \left[ \frac{1}{1} \Sigma^{-1} + \frac{1}{1} \Psi^{(i)} \Psi^{(i)} \right]^{-1} \tag{21}$$

and the matrix $\Psi^{(i)}$ is defined in (7). Assuming the channel variations follow an AR model of order 1, elements of (22) can be calculated as given Appendix 1.

Note that $K \Xi \times 1$ vector $\mu^{(i)}$ and $K \Xi \times K \Xi$ matrix $\Sigma^{(i)}$ in (21) may be expressed as in terms of subblock vectors and matrices corresponding to each user as

$$\mu^{(i)} = \left[ \left( \mu_1^{(i)} \right)^T, \left( \mu_2^{(i)} \right)^T, \ldots, \left( \mu_K^{(i)} \right)^T \right]^T,$$

$$\Sigma^{(i)} = \left[ \Sigma_{k,l}^{(i)} \right]_{k,l=1}^{K} \tag{22}$$

where $\mu_k^{(i)} = \mu_k^{(i)}[q], q = 1, 2, \ldots, \Xi$, and $\Sigma_{k,l}^{(i)} = \Sigma_{k,l}^{(i)}[p, q], p, q = 1, 2, \ldots, \Xi$.

Now let us compute the expectations on the right hand side of Eq. (18). For notational simplicity define $q \triangleq \lceil \frac{m}{L} \rceil + 1$, $m = 1, 2, \ldots, L = \Xi$.

$$a_k^{(i)}(q) \triangleq E(a_k(q)|y, b^{(i)}) = \mu_k^{(i)}[q]$$

$$\| a_k^{(i)}(q) \|^2 \triangleq E(\| a_k^{(i)}(q) \|^2) y, b^{(i)}$$

$$\triangleq \Sigma_{k,k}^{(i)}[q, q] + \mu_k^{(i)}[q] \mu_k^{(i)}[q]^*$$

$$(a_k^{(i)}(q) a_j(q))^{(i)} \triangleq E(\{ a_k^{(i)}(q) a_j(q) \}|y, b^{(i)})$$

$$= \Sigma_{k,j}^{(i)}[q, q] + \mu_j^{(i)}[q] \mu_k^{(i)}[q]^*. \tag{23}$$

Maximization-Step (M-Step): From (11) and (12), the second step to implement the EM algorithm is the M-Step where the parameter $b$ is updated at the $(i+1)$th iteration according to

$$b^{(i+1)}_k = \arg \max_b Q(b|b_i) = \sum_{k=1}^{K} \frac{Q_k(b_k)}{b_k} \tag{24}$$

M-Step can be performed by maximizing the terms $Q_k(b_k|b^{(i)})$ individually in (24), as follows

$$b^{(i+1)}_k = \arg \max_{b_k} Q_k(b_k|b^{(i)}) \tag{25}$$

where from (15)

$$Q_k(b_k|b^{(i)}) = \sum_{m=0}^{M-1} b_k(m) \Re\{ f_k^T ((a_k^+(\lceil \frac{m}{L} \rceil + 1)) x_k(m))^{(i)} \}. \tag{26}$$

Moreover, when no coding is used, since $b_k(m)$ are independent of each other, it follows from (25) that each component of $b^{(i+1)}_k$ can be separately obtained by maximizing the corresponding summation in the right-hand expression, as follows

$$b^{(i+1)}_k(m) = \text{sgn} \left[ \Re\{ f_k^T ((a_k^+(\lceil \frac{m}{L} \rceil + 1)) x_k(m))^{(i)} \} \right] \tag{27}$$

where the term $(a_k^+(\lceil \frac{m}{L} \rceil + 1)) x_k(m))^{(i)}$ has been previously obtained in (18) and $\text{sgn}(.)$ denotes the signum function. Note that it was shown in [12] that for large observations frame $M$, the first term in (23) is negligible compared to the second one. By using this property and substituting (18) into (27), along with (23) we obtain

$$b^{(i+1)}_k(m) = \text{sgn} \left[ \Re\{ f_k^T ((a_k^+(\lceil \frac{m}{L} \rceil + 1)) x_k(m))^{(i)} \} \right]$$

$$+ \beta_k (a_k^+(\lceil \frac{m}{L} \rceil + 1)^{\theta}) + \sum_{j=1, j\neq k}^{K} \rho_{kj} b^{(i)}_j(m) \left( a_j(\lceil \frac{m}{L} \rceil + 1)^{\theta} \right) \right\} \right] \tag{28}$$

where $z_k(m) = f_k^T y(m)$ and $a_k^+(\lceil \frac{m}{L} \rceil + 1)^{\theta}$ shows estimate of the time varying channel. Estimate of channel is obtained by (22) after initialization which is explained in the next section.

As a conclusion, Equation (28) can be interpreted as joint channel estimation and data detection with partial interference cancelation. At each iteration step during data detection, the interference reduced signal is fed into a single user receiver consisting of a conventional coherent detector. As a result, a $K$-user optimization problem has been decomposed into $K$ independent optimization problems which can be resolved in a computationally feasible way. Finally we remark that this paper is an extension of the work [12] to the problem of joint channel estimation and data detection for the uplink multicarrier DS-CDMA systems operating in the presence of the time -varying channels. In [12] the same problem is investigated for DS-CDMA systems in the presence of flat fading channels.
A. Initialization

The initial under sampled MMSE channel estimation has been evaluated based on pilot symbols. The corresponding pilot symbols positions, \( p_1, \ldots, p_p \), are regularly distributed over frame length \( M \) starting with \( p_1 = 1 \) and ending with \( p_p = M \). The remaining channel coefficients are obtained using linear interpolation. Initial MMSE estimate of \( B(m) \) data bits is computed from the observation of \( z(m) \) while assuming the channel coefficients have already been estimated. We refer to this method for obtaining \( a(m) \) and \( B(m) \) as MMSE separate detection and estimation (MMSE-SDE) scheme.

B. Optimal Selection of \( \beta_k \)

In usual parameter estimation problems in the presence of superimposed signals it has been shown that the optimal values of the coefficients \( \beta_k \)’s are chosen as equal weights that is \( \beta = 1/K \) [9]. However, the equally selected weights will not be optimal when the received SNR’s of each of the \( K \) users are not equal to each other and if there is some correlations between the super imposed signals, as the case we consider in our work. The optimal \( \beta \) values can be determined in this case so as to minimize the bit-error probability as the number of iterations \( i \) goes to infinity. Since this is a mathematically intractable nonlinear optimization problem we will adopt a more manageable yet a suboptimal approach presented by Kocian and Fluery [12] and extend their method for the case when each user is affected by a different flat-fading, time-varying channel. As pointed out in [12], a tractable way to determine the optimal coefficients of all the users \( \beta = [\beta_1, \beta_2, \ldots, \beta_K]^T \) is to minimize the total linear mean-square error between the true signal components \( s_k(m) = f_k b_k(m) a_k([m/L]+1) \) and their estimated values at the \( i \)th iteration \( s_k^{i}(m) \triangleq E \{ x_k(m) | y, b^{(i)} \} \) for \( k = 1, 2, \ldots, K \), after projected on \( f_k \). That is

\[
\beta_{k,\text{opt}}^{(i)} \triangleq \arg\min_{\beta_k} E \left\{ \| f^T_k \left( s_k^{i}(m) - s_k(m) \right) \|^2 \right\}. \tag{29}
\]

From (17) it follows that

\[
s_k^{(i)}(m) = f_k b_k^{(i)}(m) a_k([m/L]) + \beta_k \left( y(m) - \sum_{j=1}^{K} f_j b_j^{(i)}(m) a_j([m/L]) \right). \tag{30}
\]

Substituting (9) in (30), assuming \( w(m) \approx 0 \) [17], and taking into account the fact that the channel is asymptotically known, that is \( a_k^{(i)}([m/L]) \rightarrow a_k([m/L]) \) as \( i \rightarrow +\infty \), the terms on the left hand side of (30) can be expressed as

\[
f_k^T s_k^{(i)}(m) = b_k^{(i)}(m) a_k([m/L]) \sum_{j=1}^{K} \rho_{kj} a_j([m/L]) \left( b_j(m) - b_j^{(i)}(m) \right) \tag{31}
\]

where \( \rho_{kj} \triangleq \frac{f_k f_j}{f_j^T f_j} \). Substituting (30) and (31) in (29) and after some algebra yields

\[
\beta_{k,\text{opt}}^{(i)}(-2\beta_k + \beta_k^2) = E \left\{ [a_k([m/L])|^2 \left| b_k(m) - b_k^{(i)}(m) \right|^2 \right\} + \beta_k^2 \sum_{j \neq k} \rho_{kj} E \left\{ [a_j([m/L])|^2 \left| b_j(m) - b_j^{(i)}(m) \right|^2 \right\}. \tag{32}
\]

The expectations above can be evaluated as follows.

\[
E \left\{ [a_k([m/L])|^2 \left| b_k(m) - b_k^{(i)}(m) \right|^2 \right\} = \sigma_k^2(q) \left[ 1 - \mathbb{R}(b_k(m)b_k^{(i)}(m)) \right] = \sigma_k^2(q)P_{b_k}^{(i)}(q) \tag{33}
\]

where \( P_{b_k}^{(i)}(q) \triangleq \text{Prob} \{ b_k^{(i)}(m) \neq b_k(m) \} \) and from (ref29), \( \sigma_k^2(q) \triangleq \Sigma_{k,k}[q, q] + \Sigma_{k}[q, [\mu_k][q]]^* \). Differentiating (33) with respect to \( \beta_k \), \( k = 1, 2, \ldots, K \), equating the resulting equations to zero and solving for \( \beta_k \)’s we have

\[
\beta_{k,\text{opt}}^{(i)} = \frac{\sigma_k^2(q)P_{b_k}^{(i)}(q)}{\sum_{j=1}^{K} (\rho_{kj}^2 P_{b_k}^{(i)}(q))}, \quad 0 \leq \beta_{k,\text{opt}}^{(i)} \leq 1. \tag{34}
\]

It can be seen from (34) that \( \beta_{k,\text{opt}}^{(i)} \) is depend on the the time index \( q = ([m/L] + 1), m = 1, 2, \ldots, M = \Xi L \), since channel coefficients are time-varying over the entire observation frame length \( M \). In the case of \( \rho_{kj} = 0 \), the interferences caused by other users can be fully eliminated and the best performance corresponding to a single user detector performance can be achieved. The bit-error probability \( P_{b_k}^{(i)}(q) \) can be evaluated by assuming the performance of the multuser detector is close to a single user detector performance.

In this case \( P_{b_k}^{(i)}(q) \approx Q(\sqrt{2(\gamma_i^{(k)}(q))^2/\sigma^2}) \) where \( Q(.) \) is the error function defined by \( Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt \).

It follows from (34) that the coefficients \( \beta_{k,\text{opt}}^{(i)} \)’s are not equal to each other as will be seen in the computer simulations, the BER performance of each user is better than the case if these coefficients were selected to be equal to each other.

IV. Simulations

In this section, performance of an uplink DS-CDMA system based on the proposed receiver operating over time-varying channels is investigated by Monte Carlo simulations. The system parameters are chosen as follows: Number of users \( K = 8 \), number of pilot symbols \( p = 4 \), frame length \( M = 40 \), \( T_s = 136 \times 10^{-6} \), and equal cross-correlation coefficients \( \rho_{k,k'} = 0.4, k, k' = 1, \ldots, 8, k \neq k' \).

We investigate here two different pilot insertion scenarios to assess performance. The EM-JDE algorithm was proposed for static channels by [12] for DS-CDMA systems. Therefore, the first scenario are comparable to those used in [12] and pilot symbols are inserted at the head of the frame. In Fig.2, to demonstrate the performance degradation of the EM-JDE algorithm employing static channel model in case of using time-varying channels, three iterations have been carried out. Bit error rate (BER) performances as function of bit energy to noise power density ratio (E_b/N_0) were obtained for \( f_a = 0 \) Hz, \( f_a = 25 \) Hz and \( f_a = 50 \) Hz Doppler shifts. It was observed that proposed EM-JDE in [12] outperforms MSE-SDE for \( f_a = 0 \) Hz.
In the second scenario, the p pilot bits are regularly spread within the frame to track the channel. In this case, the normalized time spacing between two adjacent known bit positions is equal to $p_{j+1} - p_j = 13$. According to Shannons Sampling Theorem we can expect, that the EM-JDE scheme provides accurate estimates of the channel coefficients as long as the normalized Doppler frequency does not exceed $f_dT_s = 0.5 / (p_{j+1} - p_j) = 0.038$. Therefore, the cut-off Doppler frequency for a given sampling time is $280 \text{Hz}$. In Fig.3, the proposed EM-JDE algorithm for time varying channels is compared to the EM-JDE algorithm employing static channel model for three iteration. It was shown that MSE-SDE performance is improved according to first scenario because of the usage of combtype pilot pattern. It was also demonstrated that the proposed EM-JDE algorithm is able to perform better than the EM-JDE employing static channel model. Moreover, it was demonstrated that smart combination of data detection and channel estimation in JDE outperforms the separate detection and estimation scheme. In particular, it is observed that a savings of about 2 dB is obtained at $SER = 10^{-2}$, as compared with the MSE-SDE detection.

In Fig.4 we also investigate the selection of sub-frame length ($L$) for $f_d=50\text{Hz}$ and $f_d=100\text{Hz}$. In the case of $L=4$, it was shown that $BER = 10^{-2}$ performance of the EM-JDE is degrading after 1 dB and 2 dB for $f_d=50\text{Hz}$ and $f_d=100\text{Hz}$ respectively. Therefore, it was concluded that EM-JDE must calculate all channel variations ($L = 1$) for higher SNR and Doppler frequencies.

V. CONCLUSIONS

We presented an efficient iterative receiver structures of tractable complexity for joint multiuser detection and multitichannel estimation (JDE) of direct-sequence code-division multiple-access signals in the presence of time-varying flat fading channels. The schemes result from an application of EM algorithms. The EM-JDE receiver updates both the channel parameters and the data bit sequences in parallel. A closed form expression was derived for the data detection which incorporates the channel estimation as well as the partial interference cancelation steps in the algorithm. It was concluded that few pilot symbols were sufficient to initiated the EM algorithm very effectively. A comparison with other previously known receiver structures was also made. In static channel, we observed that all schemes perform almost similarly. However, computer simulations demonstrated the effectiveness of the proposed algorithms in terms of BER performances when the channel is fast fading. We conclude that the robustness of the EM-JDE which smartly combine the data detection and channel estimation in multiuser systems, unlike architectures where both process are implemented separately.
APPENDIX A

AR MODEL CALCULATIONS

We restrict ourselves to a first-order AR model, although all the derivations presented in this work are applicable to higher order models. Assuming the channel variations follow an AR model of order 1, the true channel coefficients are related to each other by

\[ a_k(q) = \gamma_k a_k(q-1) + \epsilon_k(q), \quad q = 1, 2, \ldots, \Xi; \quad k = 1, 2, \ldots, K \]

where \( \gamma_k \) is the time correlation coefficient between observation blocks of each user. The initial values \( a_k(0), k = 1, 2, \ldots, K \) are assumed to be known. Using matrix notations, (35) can be expressed as

\[ M_k \mathbf{a}_k = \mathbf{\xi}_k + \mathbf{\epsilon}_k \]

(36)

where

\[
M_k \triangleq \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
-\gamma_k & 1 & 0 & \cdots & 0 & 0 \\
0 & -\gamma_k & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & -\gamma_k & 1
\end{bmatrix};
\]

\[
\mathbf{\xi}_k \triangleq \begin{bmatrix}
\gamma_k a_k(0), 0, 0, \ldots, 0 \end{bmatrix}^T;
\]

\[
\mathbf{\epsilon}_k \triangleq \begin{bmatrix}
\epsilon_k(1), \epsilon_k(2), \ldots, \epsilon_k(\Xi) \end{bmatrix}^T.
\]

and

\[
\mathbf{a}_k = \begin{bmatrix}
a_k(1) \\
a_k(2) \\
\vdots \\
a_k(\Xi)
\end{bmatrix}^T. \text{ Since } \epsilon_k \sim N(0, \sigma^2 \mathbf{I}_K), \text{ it follows from (36) that}
\]

\[
m_k \triangleq E(\mathbf{a}_k) = M_k^{-1} \mathbf{\xi}_k = a_k(0) [\gamma_k, \gamma_k, \ldots , \gamma_k]^T;
\]

\[
C_k \triangleq E \left\{ (\mathbf{a}_k - \mathbf{m}_k)(\mathbf{a}_k - \mathbf{m}_k)^T \right\} = \sigma^2 M_k^{-1}(M_k^{-1})^T.
\]

(37)

After some algebra the \((i,j)\)th component of \(C_k\) for \(j = 1, 2, \ldots, \Xi\) can be found as

\[
C_k(i,j) = \begin{cases}
\sigma^2 \gamma_k^{-1} & \text{if } i = 1 \\
\sigma^2 (\gamma_k^{i+j-2} + \gamma_k^{i+j-4} + \cdots + \gamma_k) & \text{if } i < j.
\end{cases}
\]

It can be shown that \(C_k^{-1} = M_k^T M_k / \sigma^2 \) for \(k = 1, 2, \ldots, \Xi\).

REFERENCES


Hakan Doğan was born in Istanbul, Turkey, on 1979. He received the B.S., M.S. and the Ph.D. degrees in electronics engineering from Istanbul University, Istanbul, Turkey, in 2001, 2003, 2007 respectively. From 2001 to 2007, he was a Research Assistant at the faculty of the Department of Electrical and Electronics Engineering, University of Istanbul, working on signal processing algorithms for wireless communication systems. In 2007, he joined the same faculty as an Assistant Professor. His general research interests cover communication theory, estimation theory, statistical signal processing, and information theory. His current research areas are focused on wireless communication concepts with specific attention to equalization and channel estimation for spread-spectrum and multicarrier (orthogonal frequency-division multiplexing) systems.

Hakan A. Çirpan (M97) received the B.S. degree from Uludag University, Bursa, Turkey, in 1989, the M.S. degree from the University of Istanbul, Istanbul, Turkey, in 1992, and the Ph.D. degree from Stevens Institute of Technology, Hoboken, NJ, in 1997, all in electrical engineering. From 1995 to 1997, he was a Research Assistant at Stevens Institute of Technology, working on signal processing algorithms for wireless communication systems. In 1997, he joined the faculty of the Department of Electrical and Electronics Engineering, University of Istanbul. His general research interests cover wireless communications, statistical signal and array processing, system identification, and estimation theory. His current research activities are focused on signal processing and communication concepts with specific attention to channel estimation and equalization algorithms for space-time coding and multicarrier (orthogonal frequency-division multiplexing) systems. Dr. Çirpan is a member of Sigma Xi. He received the Peskin Award from Stevens Institute of Technology as well as the Prof. Nazım Terzioglu Award from the Research Fund of the University of Istanbul.
High-Level Design Methodology for Ultra-Fast Software Defined Radio Prototyping on Heterogeneous Platforms

Christophe Moy and Mickaël Raulet

Abstract—The design of Software Defined Radio (SDR) equipments (terminals, base stations, etc.) is still very challenging. We propose here a design methodology for ultra-fast prototyping on heterogeneous platforms made of GPPs (General Purpose Processors), DSPs (Digital Signal Processors) and FPGAs (Field Programmable Gate Array). Lying on a component-based approach, the methodology mainly aims at automating as much as possible the design from an algorithmic validation to a multi-processing heterogeneous implementation. The proposed methodology is based on the SynDex CAD design approach, which was originally dedicated to multi-GPPs networks. We show how this was changed so that it is made appropriate with an embedded context of DSP. The implication of FPGAs is then addressed and integrated in the design approach with very little restrictions. Apart from a manual HW/SW partitioning, all other operations may be kept automatic in a heterogeneous processing context. The targeted granularity of the components, which are to be assembled in the design flow, is roughly the same size as that of a FFT, a filter or a Viterbi decoder for instance. The re-use of third party or pre-developed IPs is a basis for this design approach. Thanks to the proposed design methodology it is possible to port “ultra” fast a radio application over several platforms. In addition, the proposed design methodology is not restricted to SDR equipment design, and can be useful for any real-time embedded heterogeneous design in a prototyping context.

Index Terms—Software-defined radio, design methodology, heterogeneous platform, cross-layer design

I. INTRODUCTION

This paper proposes a method that meets most of the requirements associated with the design of Software-Defined Radio (SDR) equipments. SDR related research aims at investigating all the topics that can help improving future radio systems technologies [1], [2], [3]. In this paper, we address the particular issue of SDR equipments design, which is still a very open subject. This topic is also very similar to any real-time embedded heterogeneous design issue. SDR design is indeed a co-design issue, which is not restricted to SDR, and concerns most of the embedded real-time equipments. If we could discriminate SDR from other equipment categories, we could say that SDR brings the flexibility paradigm to its height. That is why SDR equipments are expected to be made of various flexible processing components, such as DSPs, GPPs, FPGAs and ASICs. Moreover, the flexibility is not only considered at design time in SDR systems, but also at running time. An SDR system consequently should be capable of taking benefit from the potential reconfigurability of the flexible components it is made of. This implies an appropriate design methodology, which transforms this potential reconfigurability into an effective flexibility at run-time [4].

The existence of a reconfiguration management architecture at equipment level, and at infrastructure system level [5] must not be forgotten. However reconfiguration management is not directly included in the design methodology in this paper, but we explain how this approach is compatible with its insertion, in a future step. This assertion relies on our deep experience in reconfiguration management issues [6] and [7]. Reves et al. in [8] are also investigating ways to resolve these problems in a similar spirit.

There are several ways to consider SDR design. The approaches can be divided into two distinct categories:
- fast prototyping for lab demos,
- commercial equipments design.

Time has not come yet to consider the second category as a realistic option. Only sub-parts of the design flow are achieved and they are mainly those used in a prototyping flow. We particularly address in this paper the fast prototyping issue from algorithmic simulation to multi-processing heterogeneous implementation. Usual heterogeneous embedded design approaches mix several independent tools dedicated to hardware or software, which require long efforts and can cause errors. In order to tackle the issue, we act at a higher level so that automation and acceleration are implied in the design process.

The paper is organized as follows. The requirements of heterogeneous co-design, as addressed in the SDR area, are explained in section II. But there are many ways to address the concerned co-design issue. This is explained in section III of this paper and section IV extracts the main features of a realistic flow. Section V suggests a possible instantiation of this flow and illustrates how the suggested heterogeneous approach meets section III requirements. Finally, SDR design examples based on the proposed approach are given in VI, before drawing some conclusions.

II. SDR DESIGN ISSUES

SDR design is very challenging and there is not any single solution which covers all radio engineers’ expectations. This section intends to list the main issues to be resolved. It
can be noted that other embedded systems share the same requirements: this does not restrict our approach to SDR. Advanced image processing, for example, is completely within the scope of the suggested methodology.

The ideal SDR design methodology should offer a set of characteristics facilitating the design of systems, which in turn features the following main characteristics:

- flexibility,
- multi-processing,
- heterogeneous processing,
- HW-SW co-design,
- embedded constraints,
- signal-processing simulation,
- hardware abstraction.

Moreover, portability and re-usability are constant concerns, as well as the elevation of abstraction, in order to simplify the global design process. This list is not exhaustive, but it is sufficiently challenging to be considered a very ambitious target.

A. Flexibility

The salient feature to be considered first regarding SDR is flexibility. All the previously listed elements have to be considered in light of this keyword. Radio design has been based for more than a hundred years on analog electronic components, which imposes a pre-defined transmission scheme. SDR is the answer to this limitation thanks to the recent technology progresses.

Flexibility is achieved through a digital approach. Radio applications become digital (ideally, software) and can be played in any hardware platform. The radio application (software here) consequently can be changed, just like a piece of software that can be changed on a computer. This is the end of a century of fixed-behavior radio. This is the beginning of the software radio era.

B. Multi-processing

SDR is defined as an underdeveloped version of the ideal Software Radio of Fig. 1 [1]. Software Radio comprises very demanding digital signal processing capabilities, since it brings both digital-to-analog (DA) conversion at the transmission side, and analog to digital (AD) conversion at the reception side, closer to the antenna. Consequently, multi-processing is an SDR intrinsic requirement, and its purpose is to speed up the required computations.

Software radio is currently not realistic except for low carrier frequency transmission systems. More often, at least a radio frequency (RF) translation to intermediate frequency (IF) or baseband (BB) is done in the analog mode, and this between the antenna and the AD or DA conversion. This pragmatic approach is illustrated in Fig. 2 and named Software-Defined Radio or SDR.

Even if this approach reduces the efforts supported by digital processing components, this situation remains very challenging. The constraints are so strong that the direct approach shown in Fig. 2 may fail and imposes sometimes to investigate signal processing alternative solutions in the most demanding contexts, as for UWB for instance in [9] and [10]. Nevertheless, in the most relaxed situation of a baseband digitization, digital operations are timed at a multiple of the signal bandwidth, which can then reach tens of MHz in a single band terminal for UMTS, and hundreds of MHz for multi-band base-stations. A multiple of GHz will soon be required for multi-standard SDR base-stations or UWB transmissions. Most demanding operations are those, which are over-sampled compared to the symbol rate: filtering at the transmitter exit or at the receiver entrance, as well as the very complex algorithms used for reception, such as turbo decoding.

Consequently, digital processing architectures are very complex and combine several similar devices or several kinds of processing devices. This last consideration is the topic of the next paragraph. But multi-processing, without heterogeneity, is a topic in itself. An SDR design methodology must provide some facilities, in order to map and program a distributed system. Designing an SDR equipment means that the software elements, which form the radio application, need to be allocated on the equipment hardware devices.

It has to manage communications as well as processing. It has to schedule communication and processing periods for each processing unit. This includes scheduling prediction and optimization means, in order to perform an automatic mapping and scheduling of the SW application for the HW platform. Moreover, the result of this is the demand for automatic generation of the communication glue associated with the processing operations.
C. Heterogeneous computing

The heterogeneity of devices indeed allows benefiting from the different advantages associated with each category of processing components. Let us keep in mind that flexibility is of major interest for SDR. Consequently, the most popular processing devices in SDR design are the following:

- GPPs,
- DSPs,
- FPGAs,
- digital ASICs,
- analog ASICs.

As explained earlier, analog ASICs are still necessary. SDR cannot get away from antennas, amplifiers and analog filters; this is a condition associated with the transition from the electric to the electro-magnetic world. Digital ASICs and FPGAs are able to process heavily parallel computing. ASICs are hardware solutions particularly suited for low power, very high speed computing. But they are dedicated to high quantity markets in order to compensate for the chips design cost. FPGAs provide some of the requested hardware (HW) capabilities at a lower buying and developing cost. But FPGAs offer HW flexibility, which is very important when considering SDR. Two levels of flexibility may be considered for FPGAs. Firstly, flexibility at design time, enabling to create several designs in a FPGA and run them at different periods of time. FPGAs have been used this way for years, and it is of great benefit. But FPGAs offer a new opportunity now, which consists in dynamically reconfiguring a sub-part of the FPGA gates while the rest of the gates can carry on the process. Delahaye et al. have defined and validated a design methodology, as well as flexible processing elements design approaches which are particularly interesting for SDR and involve a partial reconfiguration of FPGAs [11]. Other domains are also making investigations, as for crypto [12], but this is out of the scope of this paper.

DSPs offer the best trade-off between processing power and power consumption [13]. They are particularly appropriate in the SDR context. Manufacturers now integrate more and more co-processing capabilities dedicated to radio signal processing, such as Viterbi decoders in the TI C6416 DSP for instance. Their C language programming ability makes them really relevant for SDR. They enable portability from one DSP device to another one (with the hypothesis of good compilers).

GPPs have become a potential alternative in recent years with a tremendous increase in their computation power. GPPs have the advantage of supporting high-level programming languages, as well as highly advanced operating systems. This later feature may be helpful for portability but also for reconfiguration management purposes.

D. HW-SW co-design

The distribution of the processing elements between the different categories of processing devices needs to be done on the basis of the capabilities of each category of components, processing power, but also power consumption, heat dissipation, cost and other relevant factors must be taken into account. This topic is very complex and no automatic solution exists today. The solution relies mostly on engineers’ experience. The main challenge is to separate software (SW) from hardware (HW). Let us clarify this statement: we consider SW as being an object running a processor, and HW the object executed on a FPGA. We will retain this distinction in the rest of the article.

By HW-SW co-design, we have to restrict here to the design flow ability to integrate at a high-level, processing elements of different nature in code, such as C or VHDL, so that those are run in different processing components categories, such as GPPs, DSPs, FPGAs or ASICs. We also take into consideration the specific case of parameterizable ASIC since, if the processing is fixed, the configuration of their parameters may be done on a DSP within a C function. In this sense, an ASIC is considered as a processing unit running a dedicated functionality.

E. Object-oriented design facilities

SDR is an evolution of the radio design. Radio design comes from the electronics field, which is very much linked to the notion of hardware component. This is mainly due to technology reasons. Electronics components have been for a century the only bricks available to build a radio. On the other side, computer engineering integrated the electronics paradigm of “components” in the SW domain. This makes a component-based design approach [14] definitely suitable for SDR design demand [6]. Advantages are numerous, but let us highlight the following at least: modularity, portability, replacement by reconfiguration, re-usability, etc. The reason is that the component-based approach leads to a natural separation between signal processing and execution architecture.

The object-oriented approach indeed gives the opportunity to increase the design abstraction level. This has two main consequences. Firstly, SDR design may benefit from the latest progresses in high-level design. The use of UML for embedded systems design in general [15] and SDR design in particular [16] are some example. Secondly, we argue that the temptation of inventing a completely brand new design flow for SDR must be avoided. SDR design must be based on other design technologies in order to benefit from their advances. In that sense, we defend the following approach for SDR design. An SDR design flow has to give the opportunity to integrate processing elements (let us also call them IPs here, for Intellectual Property) made by other specialized designers or tools. Then SDR design may be seen as a kind of IPs integration process. This reduces the SDR design flow constraints supported by the processing elements or IPs design and provides the designer a higher level position, thanks to a component-based approach. This permits to consider signal processing elements as black boxes with very general characteristics, such as for instance execution time, mean power consumption, memory use, etc., and not as a succession of atomic level operations, or a set of gates. Another advantage, if not the most important, is that it allows to keep each IP design optimality, thanks to dedicated tools specific to each domain; For instance as a synthesizer for a FPGA, or as a compiler for a DSP. It is unfair to pretend being able to make a better tool than the each of these domains.
specialists. Moreover, this gives the opportunity to benefit from IPs third party developers. This announces the proliferation of IPs integrators on generic platforms in the future. The re-use of IP, moreover, is a guarantee of reliability. It also gives the opportunity to benefit from the use of IP, the content of which is protected.

F. Embedded constraints

Embedded constraints can be the following:
- hard real-time,
- memory limitation constraints,
- power consumption,
- cost,
- size.

In the SDR context, the first step to overpass is often the resolution of hard real-time data flow constraints. This often implies the consideration of memory optimization as well. These two points are state-of-the-art rapid prototyping approaches, and are of course also to be considered in commercial equipment design. A real-time guarantee is mandatory. Consequently, It may not be preferable to rely on real-time operating systems (RTOS) in that context, and when compared to a static operations scheduling.

Power consumption, as well as cost and size are not considered in a prototyping approach, but are when considering commercial systems. As a consequence, those are not in the scope of the approach proposed in this paper. A long term goal is to integrate these considerations from the very beginning of the design phase, which is currently done manually. Research studies on power consumption are numerous [2], with regards to the topic importance. Some are based on a power prediction [17] or measure [18]. Attempts to help taking into consideration several of these parameters at the very early steps of a design flow have been proposed [19], but without providing an integrated resolution method yet.

G. Signal-processing integrated simulation

Another required feature of an SDR design flow is the capability to simulate the system in order to check its functional and non functional behavior. A major point to be stressed is that there is almost no interest for such a simulation if the conformity between the simulated version and the final version installed in the real system is not guaranteed. If not, the checking will have to be done again in the platform.

An alternative option is to give the possibility to implement the system very quickly on the hardware platform, and directly run the real system in order to check it in-situ.

H. HW abstraction

Abstraction layers may have several roles in an SDR system. At design time, it may be seen as a technique that enables to abstract the signal processing IPs from the hardware target, as a Java code planned to run on a Java virtual machine. This is not very good for SDR, since it may decrease the processing performance too much. Another approach is to take benefit from a higher level design approach, in order to hide some details of the hardware implementation. This may be achieved through the use of only a few features for the characterization of IPs execution on this hardware. This permits to speed-up both design phase and direct implementation on the platform, where very precise measurements may be done with exact results.

Another abstraction is acting at run-time. This may be done thanks to a middleware layer. The most common proposal related to this, is the Software Communication Architecture (SCA) selected by the US Department of Defense (DoD) in the JTRS (Joint Tactical Radio System) program [20]. A CORBA (Common Object Request Broker Architecture) software bus is proposed in the SCA, in order to support the abstraction of the hardware for the software. This is a major source of concern for the radio design community. Solutions trying to go round the full-CORBA approach while supporting some SCA compatibility are numerous [21]. They mostly consist, for real-time signal processing, in bypassing CORBA which should definitely be restricted to reconfiguration management. As explained earlier, reconfiguration management is not within the scope of the current paper but has to be taken into account, as attempted in [6], [7], [8].

I. A design approach with a wider scope than the sole SDR domain

The list of properties for SDR design approach has many commonalities with other application fields. This means that solutions targeting SDR design may also be applicable in other domains. In the context of image processing in particular, all listed characteristics are valid, even the flexibility. This is particularly relevant regarding the very latest versions of video coding schemes from MPEG standardization groups. MPEG-4 for instance targets a large scale of coding/decoding schemes which implies flexibility by nature. This requirement is also stressed at its maximum in recent RVC (Reconfigurable Video Coding) where reconfigurability is a goal itself [22]. Another example is the need to solve massively heterogeneous multi-processing issues related to the networking infrastructure, which implement video and audio formats transcoders, in order to broadcast multimedia contents on different media types (from high definition HD TV to handheld DVB-H).

III. CURRENT AND INNOVATIVE SDR DESIGN APPROACHES

A. Today: no existing solution

Regarding the design flow requirements listed in the previous section, no integrated solution for the design of SDR systems currently exists and it may be expected that none will ever be found. Consequently, a set of solutions have to be addressed separately, and then combined. We would rather expect that a new tool shall be used, in combination with already existing tools, dedicated themselves to design subparts. We suggest here to present a summary of the potentially expected solutions, on the basis of the already existing state-of-the-art languages and commercial proposals.
B. Co-design languages

If we consider the SDR design issue as a co-design issue, let us consider first the languages that have been created in the last 10 years for that purpose. We may refer to SystemC [23], HandelC from Celoxica, and CatapultC from Mentor Graphics, to discuss the popular ones. The initial goal was to provide designers with a completely integrated flow for the SW joint design, SW being the processors coding, and HW, FPGA coding. When using the term integrated, we refer to the use of a unique language that would be compiled for the SW part of the system and synthesized for the HW part after a joint edition, simulation, debug and validation process. However, the SW and HW worlds are intrinsically different if not antagonist, which makes it very challenging. Another idea was to facilitate the HW design, while getting rid off the VHDL coding and turning it to C-like. The previously mentioned languages proposed a C-like programming method for HW, while creating C++ libraries depicting HW design requirements, such as binary and vector data types, or the possibility to express parallelism. But all C++ code can not be converted into HW, which leads to a lot of synthesis impossibilities. The restrictions are so important that we almost not exaggerate if we say that it finally consists in writing VHDL in C++, which brings no help finally. Moreover, this approach is far from providing the most efficient FPGA design in terms of processing speed, surface occupation and power consumption. This finally becomes accurate with regards co-simulation, but as the HW part has to be manually reprogrammed, this completely breaks the design flow, since the validation at a high level is not guaranteed. We must stress here on the following point: a 95% synthesis of the HW target SystemC code is not important enough, if a huge amount of energy is needed to solve the 5% remaining issues, which is often the case. Finally, these co-design languages are often restricted to transfer level simulations at their best capability.

C. MathWorks

A very popular method for fast prototyping is proposed by Mathworks, based on Simulink. The first solution was proposed in the early 2000 in association with LYRtech, an SDR platform provider [24] and has been generalized to many other providers since. It consists in providing a direct bridge between a Simulink environment for signal processing simulation and the execution on a heterogeneous hardware platform made of DSPs and FPAGAs. The link from Simulink to DSPs is obtained through RealTime Workshop from MathWorks. With regards the HW side, Xilinx is providing a set of FPGA IPs, which are guaranteed to be cycle accurate equivalent to the IPs provided in Simulink. It provides indeed an artificial translation from Simulink to the FPGA. However, the following restrictions have to be taken into account: It is firstly dependent on the existence of appropriate APIs (Application Programming Interface) for each platform. If this solution was generalized, all platform providers would have to make the effort to provide those APIs. Secondly and most importantly, the HW domain is restricted to Xilinx provider and technology. Even worse, only a sub-set of Xilinx IPs is valid in this approach. The constraint on the IPs is that an equivalent cycle accurate version of the IPs should exist in both Simulink and VHDL. Note that this does not prevent designers to make their own IP with this technique. The guarantee of the equivalence is then under the responsibility of each company.

Despite the above mentioned concern, MathWorks offers the most effective solution to SDR designer in the short-term. Just a lack of a higher modeling introductory work at specification level in the design flow may be missed. However, it really permits to implement a heterogeneous design in a Simulink environment from functional simulation to proper implementation. It is, in this sense, a signal processing based approach.

D. High-level design

Moreover with longer term objectives, we may refer to computer-science oriented approaches based on high-level modeling, such as UML (Unified Modeling Language) of the OMG (Object Management group). UML is another way to interpret co-design. As UML indeed is dedicated to help modeling any kind of system, even outside the engineering domain, then why not for mixed SW and HW? In addition, systems in general become more and more complex. Hardware and software designers (we refer to this term usual meaning here, not to the “differentiating processors from FPGA designers” meaning) have to cooperate in larger and larger projects. This induces a need for a global system design methodology. The engineering answer has been MDE for Model Driven Engineering, and the methodology, MDA for Model Driven Architecture [25]. This approach is mainly based on UML concepts. Another feature is the two steps PIM and PSM. PIM stands for Platform Independent Model and corresponds to a modeling of the application, which is done independently from any implementation consideration. Then PSM (Platform Specific Model) adds new characteristics to the model, in order to take into account the execution HW platform. MDA, which was introduced in 2000 by the OMG, is now a quite successful concept. An attempt to integrate SDR design in a UML methodology has been done in A3S project [19], [26] for the modeling and prediction of non functional characteristics of such systems. A metamodel for SDR has been established, in order to generate a UML profile named “A3S profile” [19]. A profile is a way to extend UML concepts for a specific domain. In practice, this consists in customizing a UML environment for a specific application domain (and getting rid off all the unnecessary concepts). The OMG has already standardized several metamodels for embedded real-time systems. The latest is MARTE [15]. New researches on SDR have been inspired by MARTE, such as Mopcom [27] whose aim is to define a design flow generating automatically VHDL code from high-level UML-based modeling [28].

Note that if we push SDR design towards cognitive radio, this increases once step further complexity. High-level design, through meta-modelling, is a key to deal with system complexity, such as proposed in [29].
E. Other approaches

Other design approaches for SDR have mainly oriented their solution towards a specific hardware implementation target, while offering the inherent flexibility of SDR design. One solution is to suggest an SDR customized processor. QuickSilver Inc. has designed a hardware chip and associated software development tools. The Adapt2000 ACM System Platform claims to integrate in a single IC, the Adapt2400 ACM (Adaptive Computing Machine), ASICs, DSPs, FPGAs capabilities and micro-processors [30]. The Adapt2400 comprises four distinct heterogeneous node types and a node wrapper. The InSpire Node Software Tool Set complements the Adapt2400 architecture, and abstracts away the complexity of the heterogeneous, multi-nodal, multi-tasking ACM architecture under a single unified programming model.

Another choice is to privilege brut performance (computation speed, power consumption). A promising approach in this field is the use of systems on chip (SoC) that gathers several (more or less dedicated) processing units inside a single chip. The generalization to a high number of units, as required in SDR systems, leads to networks on chip (NoC). A NoC can be differentiated from a SoC, as the number of units become so high, that communication between units themselves becomes an issue, and they then need their own processing units. A very advanced work in the SDR domain is the FAUST (Flexible Architecture of Unified System for Telecommunication) chip designed by CEA [31]. It has been originally designed to support potential 4G candidates based on MC-CDMA modulation schemes [32]. Flexibility was a necessity in order to support several ranges of parameters and to evaluate several possibilities, and to incorporate flexible schemes for 4G. This included real-time reconfiguration capabilities in a certain set of configurations, which were limited by each unit parameters possibilities. Then it turned in a second step towards a real platform for the support of multi-standards SDR systems. CEA designed a new version of the FAUST chip named MAGALI (Multi-Applications Globally Asynchronous Low-power Integrated circuit), which incorporates new and more diverse processing units. This allows MAGALI to support a large scale of current and future standards, such as WiFi, WiMax, etc. with MIMO support. Please note that, even if the set of possibilities is very high, it is limited by specific manufacturing constraints. It is specifically dedicated radio processing and could not address other domains. It also includes advanced capabilities in terms of power consumption savings, which is a definitely advanced feature in the SDR field.

Even if this seems to be in contradiction with choosing a HW target, Vanu Inc. made the choice to remain generic a hundred percent, i.e. the hardware target they selected is “the” generic processor. It is based on the MIT researches, which were carried out in the middle of the 90’s [33] for GSM base stations [34]. The main idea is to benefit from the GPPs technological improvement, according to Moore’s Law, and keep the advantage of existing high-level programming tools, which are supported by a GPP environment. The portability is then implicit, as it is an exact application of the PC concept to the radio domain. Vanu Inc. was the first company in the USA to be certified for GSM SDR base stations, and began manufacturing in 2004. But we may wonder why this technique is not so popular and why Vanu Inc. does not have a stronger market position. This indicates that this technique is probably not yet mature for the infrastructure. It will consequently be even worse for terminals as GPPs main drawback is power consumption, a key point of terminal design.

Let us mention also the tools which are very accurate for the optimized design of FPGA IPs and may be complementary to the previous ones. For example: LISAtek [35], and GAUT [36]. LISAtek aims at providing integrated tools for the design of ASIPs (Application Specific Instruction Set Processor), in order to obtain an optimized hardware processor architecture for a dedicated set of processing operations and the associated programming tools (debugger, compiler). This could be used for the design of an SDR SoC for instance, as experimented in [37]. GAUT enables to reuse an IP algorithmic specification written in C/C++, and to synthesize it into an equivalent VHDL RTL specification. The potentially pipelined architecture which is generated, allows exploring the design space for FPGAs or ASICs, through a trade-off between processing speed, power consumption and area. This can be a complementary tool for the integration of the processing units into the previous solutions.

From the system point of view, another tool worth to be considered for the design of embedded equipments is CoFluent Studio, made by CoFluent Design [38]. The tool allows to manually exploring a heterogeneous design space with an extreme intimacy, in order to anticipate and orient design possibilities. This is a product-oriented approach which requires quite some time and expertise. Moreover, CoFluent does not provide automatic code generation facilities. The exploration stays at the level of a virtual platform. We mostly privilege in this paper rapid prototyping oriented solutions, which help a signal processing engineer to implement quickly an SDR system with automated steps.

Future will tell if one of those approaches more or less dedicated to SDR, is an appropriate answer or not. Maybe the SDR market will be shared between several of them. The design methodology suggested in the next section is less dedicated to SDR than most of those described here, and could also be worth for other embedded real-time prototyping design perspectives.

IV. A REALISTIC DESIGN FLOW FOR SDR OPEN ARCHITECTURES

The term “realistic” is used here to explain that, with current state-of-the-art technologies, a perfect solution for fully integrated co-design is not available, and even not foreseen. The term “open architectures” means that we consider the solutions based on COTS programmable and reconfigurable components. We exclude from the study any approaches based on pre-oriented hardware, which reduces its range to only a subset of SDR capabilities in terms of flexibility. The solutions suggested here can be used for any heterogeneous real-time data-flow oriented embedded design.
However, the concern of this paper is SDR, so we must keep in mind that SDR is not only data-flow oriented. The approach has to be compatible with the adding of a reconfiguration management architecture at a later stage of the design, more control-oriented by definition. Moreover, SDR design must not be restricted to radio design. It should be seen from a larger point of view or scale, including other layers of the OSI (Open Systems Interconnection) model. SDR design is also a question of cross-layer design. This typically concerns, at terminal level, a joint approach for both radio and image processing for instance [39].

A. SDR Ideal design flow

Fig. 3 schematically represents what could be the ideal SDR design flow. It would, first of all, consist in a high level modelling phase. This would allow taking advantage of mature modelling techniques, which allow in turn the various design protagonists, respectively HW, SW and signal processing designers, to refine the very early design choices in a common environment issued from the system specifications.

The first step consists in making a simulation and functional validation of each IP. The hardware target that is planned to run the IPs, would be here completely transparent. Ideally, the same language would be used to program a SW (dedicated to run in a processor) or a HW (dedicated to run on a FPGA) IP.

The system functional validation is directly obtained by combining the previous step validated IPs. In conjunction with this step, non functional requirements, as well as platform features are derived. The platform has to be considered here as the combination of devices with their associated low level software as a board support package, or as RTOS, etc.

Then all the information is merged to allow an automatic HW/SW partitioning for heterogeneous multi-processing. Predictions figures are achievable, such as the application execution time, so that other HW/SW matching could be tried if the constraints are not respected. This allows dimensioning the HW platform at the early stage of the design flow. Let us just note that this is far from current reality. The partitioning guarantees the functional accuracy of the application multi-processing version, when compared to the previously simulated mono-processing version. Of course, in that ideal world, the automatic scheduling is also done. The communication code generation is derived directly from the scheduling. IPs automatic code generation is obtained, for either HW or SW. The transformation guarantees an equivalence with the previous model: this permits to avoid the repetition of the simulation and verification procedures, already done earlier.

A co-simulation tool allows to validating definitively the heterogeneous system. Finally, the implementation on the platform is straightforward as the ideal design flow takes every aspects of the implementation, whatever their level.

B. A realistic SDR design flow

A realistic SDR design flow is suggested in Fig. 4. It intends using a set of already existing commercial and/or academic technologies at their maximum capabilities.

Two main goals are achieved here. The first goal is to avoid the reprogramming of the same function several times. This is what happens usually in heterogeneous design. No less than 6 re-coding steps may be done for certain sub-parts of a heterogeneous system, as for instance a HW IP:

- Matlab functional validation,
- floating point C functional validation,
- fixed-point C validation,
- SystemC,
- cycle-accurate SystemC,
- VHDL.

In addition to the extra time necessary to code again the same functionality, this approach is very error-prone and each step requires a repetition of the debug and validation process.

The second goal consists in automating all the implementation dependent actions, and keeping the multi-processing functional accuracy from separately developed and validated IPs. The IPs typical abstraction level is C language or VHDL. This could even be an executable code or a bitstream coming from a third party, which would keep its code secret. This means that the IPs code for the selected HW target is available, or can be obtained with dedicated tools (compiler for DSP and synthesizer, place and route for FPGA), so that non functional
characteristics may be extracted, such as for instance: the programming code and data code size in a DSP, the number of gates for a FPGA, the execution speed, etc.

As previously stated, the IPs design is voluntary set out of this flow. Please note that we also definitely consider here, that there is no efficient solution for an automatic HW/SW partitioning. This must be left to the designer’s discretion, his/her choice being made on the basis of his/her experience.

Then HW (respectively FPGAs) and SW (respectively processors) worlds have to be treated separately, each of them having their own ways of managing multi-processing for automatic partitioning and scheduling. But strong equivalent principles must be applied for both; this enables to keep a global cohesion for an SDR design approach. The most important thing is the asynchronism between operations. This is the necessary base that will enable to integrate a reconfiguration management architecture at a later stage.

This gives many advantages which are also SDR requirements:
- re-usability: an IP may be used in different designs at different clock rates,
- portability: from one HW device to another,
- managing reconfiguration in a future step,
- power consumption considerations, since GALS originate from those.

Nevertheless automatic code generation for communications between both worlds has to be possible at least.

The methodology should also bring some automatic partitioning and scheduling concerning the SW side only. It should be able to provide a multi-processing version, which is a functionally certified equivalent of the application used for the mono-processing version validation.

Section V proposes a way to implement the design flow proposed in Fig. 4.

C. How can reconfiguration be supported at a later stage?

Reconfigurability is intrinsically provided within this methodology by two complementary means:
- the choice of generic hardware components for the platform, such as GPPs, DSPs, FPGAs,
- through the software application building, which is done via a component-based approach for both SW (processors) and HW (FPGAs) processing elements.

This is the necessary base that will enable to integrate a reconfiguration management architecture at a later stage.

We have derived a reconfiguration management architecture that is particularly suited to this design methodology [6], [7]. It is out of the scope of this paper to describe it, but we precise that it has been successfully developed and experimented through several prototyping showcases. One main feature of an appropriate reconfiguration management architecture, is to perform reconfiguration operations in a very short time, in order to limit as much as possible the overhead, when compared to the signal processing duration.

It is important to point out that we deliberately disjoined the reconfiguration management design on one hand, and the SDR signal processing on the other hand, while keeping all the necessary interconnections between them.

We do not think that this technology is mature enough yet, so that both approaches can be merged. That is the reason why the design methodology of this paper does not include, but supports the coherence with a reconfiguration management architecture. However we are attempting to pursue this final goal, such as for the FPGA sub-part in the Mopcom collaborative project [27], [28].

Adding to this, it has been highlighted that a particular effort has been done to completely master each part of the design. We insist that this is a key point of an efficient reconfiguration process, i.e. integrate reconfiguration in the processing elements design itself. From a formal point of view, we have been investigating parameterization techniques for several years. This consists in designing processing elements in such a way, that they may be reconfigurable quickly enough through the use of parameters [41]. If we extend the scope of this approach, we can plan to use it for multi-standard equipments design in combination with the use of common operators, as explained in [42], [43], [44].

From the experimentation point of view, we have also shown that the reconfiguration management may be merged, at least partially, with the signal processing in order to offer more flexibility. That is why we propose hierarchical and distributed management architecture in [45]. It is particularly obvious
in the particular case of FPGAs partial reconfiguration, as investigated in [7], [11].

The transition from one configuration to another can be extracted from the difference existing between two designs, which are generated rapidly by the methodology proposed in this paper. Some overhead has to be planned for the transition from one design to another in terms of processing time.

D. Which already existing solutions fit or do not?

The combination of Simulink with platforms such as LYRTech seems to us the best solution. But there is an eliminating limit: designing one’s own IP is a necessity in order to control the system in such a sensible implementation domain. With the Simulink approach, the designer depends on Xilinx IPs to fully use the design flow pertinence. It is also possible to add your own IP, but it then requires a second coding: both for Simulink and VHDL. Then you miss one of the goals, which is to avoid re-coding.

High level design approaches based on UML offer promising perspectives to formalize the top of the design flow. A bridge is needed between specification and functional validation. A3S approach and A3S metamodel [16] partly answered this need, but without generating, as a result of the architectural study, the code for both validation and implementation. The Mopcom project is trying to fill this gap for the HW side, while using automatic transformations between three layers of metamodels [27]. Each of these layers takes into account, with an increasing level of accuracy, the implementation details. The aim is to model a system at several levels of abstractions, while keeping advantage at a lower level of the validation and results already obtained at a higher level.

We do not pretend that we have explored all the existing solutions, but we mentioned the most interesting ones. In a nutshell, we can draw the following analysis: there are two ways of considering the extension of the design flow towards a fully integrated and automatic design flow for SDR. Either starting from the electronics point of view and follow a bottom-up way. This was chosen by designers who want to keep real-time a priority, for instance. Or it is also possible to look at it in a top-down perspective and privilege a software engineering approach, in order to take full benefit from the latest computer science advances and tools.

V. A DESIGN METHODOLOGY FOR ULTRA-FAST SDR PROTOTYPING BASED ON SYNDEX

We suggest here in detail a design methodology for SDR prototyping, respecting 4 and fulfilling the requirements of part II. The set of tools we are using for that purpose is probably not the only possibility. That is why we firstly presented this methodology from a general perspective in section IV-B, and from a possible instantiation perspective in the present section. Designers are then free to use any other implementation of this methodology, according to their particular application domain or habits. We remind also that the granularity level of the constitutive elements considered in the methodology is the size of a signal processing IP, such as a filter, a Viterbi decoder or a FFT for instance.

The transition from one configuration to another can be extracted from the difference existing between two designs, which are generated rapidly by the methodology proposed in this paper. Some overhead has to be planned for the transition from one design to another in terms of processing time.

D. Which already existing solutions fit or do not?

The combination of Simulink with platforms such as LYRTech seems to us the best solution. But there is an eliminating limit: designing one’s own IP is a necessity in order to control the system in such a sensible implementation domain. With the Simulink approach, the designer depends on Xilinx IPs to fully use the design flow pertinence. It is also possible to add your own IP, but it then requires a second coding: both for Simulink and VHDL. Then you miss one of the goals, which is to avoid re-coding.

High level design approaches based on UML offer promising perspectives to formalize the top of the design flow. A bridge is needed between specification and functional validation. A3S approach and A3S metamodel [16] partly answered this need, but without generating, as a result of the architectural study, the code for both validation and implementation. The Mopcom project is trying to fill this gap for the HW side, while using automatic transformations between three layers of metamodels [27]. Each of these layers takes into account, with an increasing level of accuracy, the implementation details. The aim is to model a system at several levels of abstractions, while keeping advantage at a lower level of the validation and results already obtained at a higher level.

We do not pretend that we have explored all the existing solutions, but we mentioned the most interesting ones. In a nutshell, we can draw the following analysis: there are two ways of considering the extension of the design flow towards a fully integrated and automatic design flow for SDR. Either starting from the electronics point of view and follow a bottom-up way. This was chosen by designers who want to keep real-time a priority, for instance. Or it is also possible to look at it in a top-down perspective and privilege a software engineering approach, in order to take full benefit from the latest computer science advances and tools.

V. A DESIGN METHODOLOGY FOR ULTRA-FAST SDR PROTOTYPING BASED ON SYNDEX

We suggest here in detail a design methodology for SDR prototyping, respecting 4 and fulfilling the requirements of part II. The set of tools we are using for that purpose is probably not the only possibility. That is why we firstly presented this methodology from a general perspective in section IV-B, and from a possible instantiation perspective in the present section. Designers are then free to use any other implementation of this methodology, according to their particular application domain or habits. We remind also that the granularity level of the constitutive elements considered in the methodology is the size of a signal processing IP, such as a filter, a Viterbi decoder or a FFT for instance.

The transition from one configuration to another can be extracted from the difference existing between two designs, which are generated rapidly by the methodology proposed in this paper. Some overhead has to be planned for the transition from one design to another in terms of processing time.

D. Which already existing solutions fit or do not?

The combination of Simulink with platforms such as LYRTech seems to us the best solution. But there is an eliminating limit: designing one’s own IP is a necessity in order to control the system in such a sensible implementation domain. With the Simulink approach, the designer depends on Xilinx IPs to fully use the design flow pertinence. It is also possible to add your own IP, but it then requires a second coding: both for Simulink and VHDL. Then you miss one of the goals, which is to avoid re-coding.

High level design approaches based on UML offer promising perspectives to formalize the top of the design flow. A bridge is needed between specification and functional validation. A3S approach and A3S metamodel [16] partly answered this need, but without generating, as a result of the architectural study, the code for both validation and implementation. The Mopcom project is trying to fill this gap for the HW side, while using automatic transformations between three layers of metamodels [27]. Each of these layers takes into account, with an increasing level of accuracy, the implementation details. The aim is to model a system at several levels of abstractions, while keeping advantage at a lower level of the validation and results already obtained at a higher level.

We do not pretend that we have explored all the existing solutions, but we mentioned the most interesting ones. In a nutshell, we can draw the following analysis: there are two ways of considering the extension of the design flow towards a fully integrated and automatic design flow for SDR. Either starting from the electronics point of view and follow a bottom-up way. This was chosen by designers who want to keep real-time a priority, for instance. Or it is also possible to look at it in a top-down perspective and privilege a software engineering approach, in order to take full benefit from the latest computer science advances and tools.

V. A DESIGN METHODOLOGY FOR ULTRA-FAST SDR PROTOTYPING BASED ON SYNDEX

We suggest here in detail a design methodology for SDR prototyping, respecting 4 and fulfilling the requirements of part II. The set of tools we are using for that purpose is probably not the only possibility. That is why we firstly presented this methodology from a general perspective in section IV-B, and from a possible instantiation perspective in the present section. Designers are then free to use any other implementation of this methodology, according to their particular application domain or habits. We remind also that the granularity level of the constitutive elements considered in the methodology is the size of a signal processing IP, such as a filter, a Viterbi decoder or a FFT for instance.

The transition from one configuration to another can be extracted from the difference existing between two designs, which are generated rapidly by the methodology proposed in this paper. Some overhead has to be planned for the transition from one design to another in terms of processing time.

D. Which already existing solutions fit or do not?

The combination of Simulink with platforms such as LYRTech seems to us the best solution. But there is an eliminating limit: designing one’s own IP is a necessity in order to control the system in such a sensible implementation domain. With the Simulink approach, the designer depends on Xilinx IPs to fully use the design flow pertinence. It is also possible to add your own IP, but it then requires a second coding: both for Simulink and VHDL. Then you miss one of the goals, which is to avoid re-coding.

High level design approaches based on UML offer promising perspectives to formalize the top of the design flow. A bridge is needed between specification and functional validation. A3S approach and A3S metamodel [16] partly answered this need, but without generating, as a result of the architectural study, the code for both validation and implementation. The Mopcom project is trying to fill this gap for the HW side, while using automatic transformations between three layers of metamodels [27]. Each of these layers takes into account, with an increasing level of accuracy, the implementation details. The aim is to model a system at several levels of abstractions, while keeping advantage at a lower level of the validation and results already obtained at a higher level.

We do not pretend that we have explored all the existing solutions, but we mentioned the most interesting ones. In a nutshell, we can draw the following analysis: there are two ways of considering the extension of the design flow towards a fully integrated and automatic design flow for SDR. Either starting from the electronics point of view and follow a bottom-up way. This was chosen by designers who want to keep real-time a priority, for instance. Or it is also possible to look at it in a top-down perspective and privilege a software engineering approach, in order to take full benefit from the latest computer science advances and tools.

V. A DESIGN METHODOLOGY FOR ULTRA-FAST SDR PROTOTYPING BASED ON SYNDEX

We suggest here in detail a design methodology for SDR prototyping, respecting 4 and fulfilling the requirements of part II. The set of tools we are using for that purpose is probably not the only possibility. That is why we firstly presented this methodology from a general perspective in section IV-B, and from a possible instantiation perspective in the present section. Designers are then free to use any other implementation of this methodology, according to their particular application domain or habits. We remind also that the granularity level of the constitutive elements considered in the methodology is the size of a signal processing IP, such as a filter, a Viterbi decoder or a FFT for instance.

A. SynDEx approach

The tool selected for the design flow has been SynDEx [46], created by INRIA. SynDEx stands for Synchronized Distributed Executives. The role of this tool in the proposed methodology is to proceed the coarse grain design steps until the code generation [47]. This mainly concerns the communication code glue between IPs.

Then specialized tools dedicated to each processing domain (respectively HW/FPGAs and SWprocessors) take over to the implementation. A major requirement of the design flow is that no re-writing of the code is necessary between these two steps in order to keep the former step verifications valid.

a) Major features: Entries for the design flow are two graphs:

- a hardware platform diagram,
- a software application diagram,

In order to speed-up the design phase, it is deliberately chosen to have only a restricted set of parameters and to take into account the architectural exploration:

- execution time of each IP,
- atomic communication time for each data type,
- communication media features

The goal is to obtain an implementation on the HW platform as quickly as possible, in order to deal with realistic constraints instead of having approximated simulations of those constraints in the design flow itself.

Partitioning optimization of SynDEx is performed using graph theory. The SW application graph is modeled by an extended Data Flow Graph (DFG), which is an oriented hyper-graph. Each vertex corresponds to an algorithm operation or a processing element, which is activated by the fullness of its input buffer and each edge represents a data transfer between operations. An example of a SynDEx software application graph is given in Fig. 5.

Moreover, the model includes hierarchy, delay, conditioning (if / then / else) and factorization (for loop), in order to express the potential parallelism. Factorization, which is associated

![Graph Example](image-url)
with a repetition factor, is used to repeat operations and requires additional specific vertexes in the DFG:

1) Fork/Join vertexes: the “Fork” vertex explodes each element of the data it receives for each parallel repetition of the consumer operation, whereas the “Joint” vertex builds the data it sends, via the concatenation of each separated element produced by each producer operation repetition.

2) Iterate vertex: this vertex aims at sequentially duplicating a producer/consumer operation, where the outputs of the current operation can be the input of the next one, if the data names are similar. More details can be found in [48].

The hardware architecture is modeled by a non-oriented hyper-graph, where each vertex is a processor (hardware component) and each hyper-edge represents a communication medium, as shown in Fig. 6. In this model, a processor consists of one operator and as many communicators as there are connected media. An operator executes the operations, which are a part of the algorithm, and a communicator executes a communication operation, when a data transfer is required. By doing this, a multi-component architecture can be represented by a network of Finite State Machines (FSM) interconnected with communication media (FIFO, shared memories etc.).

The aim of this tool is to find the best combination of an algorithm which specifies the running of the application, and a multi-component architecture. In addition to this, real-time and embedding constraints must be satisfied. This methodology is based on a graph theory, in order to model the software application and the hardware architecture. The software and hardware are described by two distinct graphs. SynDEx transforms those two graphs with graph transformations in order to generate an optimized code implementation.

An efficient implementation graph is obtained thanks to optimizations and simultaneous distribution and scheduling operations, while trying to minimize the execution time.

There are a large number of possible implementations. The optimization problem aims at selecting the most efficient one between all of them (best latency). The latency is the total DFG execution time on a given HW architecture, between the first scheduled operation and the last one. Moreover, the distribution and scheduling problem in the case of HW multi-component is known to be NP-hard (an exhaustive research on all the possible fulfillments is inconceivable). This is why heuristics are used to find the the optimal solution best approximation (greedy and genetic algorithm). The current heuristic approach attempts to minimize the total running algorithm execution time on the multi-component architecture. Moreover, since the Synchronized Distributed Executives (SynDEs) are automatically generated and safe, this consequently eliminates some of the tests and low-level hand-coding, and decrease the lifecycle development as well.

SynDEx provides a timing graph, as in Fig. 7, which includes simulation results of the distributed application and thus enables SynDEx to be used as a virtual prototyping tool.

SynDEx is also providing a static scheduling, which is of major importance for hard real-time requirements. It gives the guarantee of an execution within a restricted given latency period. SDR equipment will have very strong real-time constraints to respect. The introduction of SW in radio design must not cause customer’s suspicion, or even worse, them rejecting it. This has already been witnessed in the mobile phone area with the collapse of WAP (Wireless Application Protocol), illustrating the fact that a badly introduced technological advance can turn into an economic disappointment.

b) SynDEx tool: SynDEx is a CAD tool whose original aim is to parallelize the application processing for a multi-processor network [49]. It provides a multi-processing version of an application that is functionally accurate, when compared to the mono-processor initial version. SynDEx performs an optimized partitioning and scheduling of an application used with a GPPs multi-processors architecture. Typically, the communication media is a TCP/IP in this context. This can be easily implemented in a platform, instead of a network perspective, by varying communication means between the processors, such as FIFO, dual-port memory, PCI bus, etc....

Moreover, SynDEx has been extended towards the embedded reality and in particular in terms of memory optimization [48], [50]. In this sense, SynDEx answers particularly well SDR issues in the restricted SW domain, in other words if the SDR equipment is only made of processors. SynDEx indeed is close to the ideal software radio design. Nevertheless, other SynDEx intrinsic features are worth to be used, beyond the purely SW application domain. After the multi-processor matching, SynDEx generates a scheduled version of the application for each of the processors involved in the platform. This code is generic and not dedicated to any implementation. It is written in M4 macro-code. Therefore a M4 macro-code file is obtained for each processing node of the graph describing both scheduling and communication synchronisms. Those M4 files can be translated in any programming language (assembly, C, VHDL, etc.). The translation is operated with a GNU M4 macro-translator, thanks to the use of translation...
TABLE I
LIBRARIES DEVELOPED FOR THE SUPPORT OF MULTI-TARGETS, MULTI-PLATFORMS DEVELOPED BY IETR/INSA

<table>
<thead>
<tr>
<th>Platforms</th>
<th>Processors</th>
<th>Communication media</th>
<th>Media Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sundance SMT310q</td>
<td>1MS320C64x</td>
<td>SDB (Sundance)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Sundance SMT320</td>
<td>1MS320C62x</td>
<td>SHB (Sundance)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Sundance SMT361</td>
<td>DM642</td>
<td>Comport (Sundance)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Sundance SMT395</td>
<td>x86 (windows)</td>
<td>PCI_RAM (Sundance)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Sundance SMT365</td>
<td>FPGA</td>
<td>PCI_SAM (Sundance)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Sundance SMT319</td>
<td>TCP (windows,</td>
<td>Converter ADC (Pentek)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Pentek p4292</td>
<td>C62x, C64x)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitec VP3-PMC</td>
<td>Converter DAC</td>
<td></td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>Sundance SMT348</td>
<td>VPI_RAM (Vitec)</td>
<td></td>
<td>FIFO/RAM</td>
</tr>
<tr>
<td></td>
<td>Bi-FIFO (Pentek)</td>
<td></td>
<td>FIFO/SAM</td>
</tr>
</tbody>
</table>

TABLE II
ALREADY EXISTING LIBRARIES

<table>
<thead>
<tr>
<th>Processors</th>
<th>Communication media</th>
<th>Media Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel x86 (linux)</td>
<td>TCP (linux)</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>ADSP21060</td>
<td>RS232</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>TMS320C40</td>
<td>CAN</td>
<td>FIFO/SAM</td>
</tr>
<tr>
<td>MC68332</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPC555</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It also efficiently supports devices which are controlled by processors, such as analog or digital ASICs. And, in order to adapt it to SDR design, it is mandatory to extend it to reconfigurable hardware computing, namely FPGA.

HW implication has indeed to be considered outside the SynDEx partitioning optimization, if we want to keep HW efficiency (of parallelism), since the processor world is sequential. That is why we mentioned earlier that HW/SW partitioning was not supported by this methodology. The risk related to the HW and SW approaches in a single automatic methodology, is to decrease one side performance for the benefit of the other side. Optimization purposes may be indeed at opposite ends, with regards HW and SW. This can be considered as a granularity issue, between respectively the description of an efficient algorithm written in SW (coarse grain), C language, or HW (fine grain) VHDL.

Therefore, we suggest a methodology which integrates the HW design with the SW and system design methodology presented till now, while the HW/SW partitioning has been manually done. The characteristics that should be kept in the HW side and that have to be fully compatible with the design at the SW side, are the following:

- based on a component-based approach,
- HW portability from one design to another, whatever the device, the clock, etc.,
- HW to SW IPs migration,
- the support of IPs made of gates as well as IPs running on embedded processor cores inside FPGAs,
- the support of ASIC.

c) FPGA support: A reformulation of the design approach has to be made here. The way to solve the interaction issue between an intrinsically sequential and a highly parallel approach lies in the response to the following question: How should we reformulate the data-flow approach related to the processors sequential world within the reconfigurable hardware parallel world, without loosing parallel HW highly efficient processing speed? Let us just recall that one major goal in SDR is to add, delete, replace a processing element, without causing any disturbance to the other processing elements of the radio design being inside or outside the FPGA.

The first part of the solution is solved thanks to the IP approach, which guarantees the parallelism inside the IPs, since it has been designed with usual HW tools. In other words, SynDEx allows keeping the execution parallelism inside an HW IP with this approach, whereas SynDEx is designed so that it executes sequential operations inside a single processor, and only extracts parallelism between several IPs on a multi-processing architecture.

The second part of the solution uses GALS, as already addressed in section IV-B. That permits to provide also parallelism through pipeline. GALS permits to extend the IPs processing asynchronism to the HW world, independently from the data flow rhythm (just has to be faster). Input buffers, associated with adequate hand shaking signals of Fig. 8 permit to launch and stop the IPs operation, depending on the presence (or not) of data at the input of the IP, or at the place available in the buffers at its output.

libraries, considered as dictionaries. We have been developing translation libraries for many different embedded targets, as shown in Table I, added to the already existing ones shown in II.

Libraries for platforms, processors and communication media are listed in both tables. Some libraries depend only upon the language (C, VHDL), others depend on the processor nature, and finally on other devices, typically those used for communications (FIFO, memory, bus, etc.).

Since SynDEx is dedicated to the processor world (called SW in this paper), it extracts parallel operations from the application graph, in order to map them on different processors (hardware components). However, it performs IP sequential operations on each of the HW graph processors. This consequently does not match with the HW (FPGA) reality, which benefits from parallel execution. However all the code generation and scheduling carried out in the SW world on the one hand, and between the HW and SW world on the other hand can be kept in totality. The HW domain content has only to be considered separately, bearing in mind that common principles are shared with the SW side.

B. Heterogeneous processing support

This methodology is originally dedicated to multi-processor (of any kind: DSP, GPP, µC, etc.) architectures programming.
Fig. 8. GALS schematic encapsulation.

Fig. 9. Data flow speed adaptation in a GALS if $f_1 > f_2 > f_3$.

GALS has three main interesting features in our methodology, as illustrated in Fig. 9:
- It adapts the IP operation average frequency to the other IPs frequencies of the chain (9b for IP A),
- This allows to change an IP of clock domain (9c),
- This allows blocking an IP from time to time, as for a reconfiguration operation for instance (9d).

d) Specific design use cases with FPGA: In the specific case of a unique IP running on a FPGA, the methodology is one hundred percent useful and automatic, as there is no sequential issue for one IP. The FPGA is then a HW accelerator. In this particular context, an automatic HW/SW partitioning is even possible as illustrated in Fig. 11. Let us consider the example of the software application in Fig. 10. If one has to decide whether IP4 should be inside a FPGA or a DSP, then the condition is to have both code versions, and their corresponding execution time for both targets. Then, the methodology is able to deal with such a context and may optimize and generate the full code for both DSP and FPGA with the scheduling obtained on Fig. 11.

Fig. 10. Software application graph example.

The methodology takes the decision to instantiate IP_4 on the FPGA instead of the DSP, since executing IP_4 in parallel of IP_3 and IP_5 executions, saves time: IP_4 execution time and communication overheads are completely masked during IP_5 and IP_3 executions. The FPGA acts here as a co-

processor.

A sub-optimal but entirely automatic way of using the suggested methodology is also possible with several IPs on a FPGA. As shown in Fig. 12, it consists in separating a FPGA in as many HW components as there are IPs. If IP3 and IP4 of Fig. 10 are planned to be implemented on a single FPGA, we may define two Virtual FPGAs, assigning IP3 on virtual FPGA1 and IP4 on virtual FPGA2. Fig. 12 shows how the methodology is able to parallelize IP3 and IP4 and take it into account in the global scheduling of the application. The merging of all generated VHDL code for each HW components, only needs to be done hereafter.

e) Digital and analog ASIC support: SDR design may not only involve fully programmable or reconfigurable devices, it is necessary to support also specific devices such as ASIC for instance. For example, in an SDR design, typical ASIC devices are: digital down and up converters (DDC and DUC), analog to digital and digital to analog converters (ADC and DAC), amplifiers, analog filters, antenna, etc. such as illustrated in Fig. 13. These circuit boards have to be inserted in the design flow for two reasons: firstly, they may be the destination or the source of the data to be processed. Secondly, they may be parameterizable and need to be configured through primitive functions, which are activated by a processor for instance. Therefore, the processor needs to have the necessary code necessary, so that it can perform both initialization and adjustment of these parameters. Please note that in some contexts, a DSP may change registers inside a FPGA, from which they configure the ASIC.
Fig. 12. Scheduling graph with one DSP and one FPGA for each IP.

Fig. 13. SynDEx hardware graph with ASIC (DDC, DUC, ADC, DAC).

C. A non functional time-based approach

The set of non functional characteristics to be taken into account would ideally be as many as possible, but in order to keep it feasible, it is restricted here to time. This means that each IP has only to be characterized in time in order to be integrated in the design flow. Time includes both processing and communication time. Either it is extracted from measurement, or it can be estimated. Then SynDEx processes the global application timing optimization by mapping the IPs on the hardware architecture. The precision of the global design prediction depends on the accuracy on each IP. The partitioning algorithm used in SynDEx is a greedy algorithm, which provides a good compromise between partitioning execution time and quality result. Another approach has also been developed on the basis of a genetic algorithm, in order to refine the SynDEx greedy algorithm solution [47]. This approach provides the same functional results. If the genetic algorithm is used, the cost function decreases the latency in some cases by 50%. The price is an extended computation time of many thousands compared to a solution obtained by a SynDEx greedy solution.

Moreover, on top of time considerations, some memory considerations may be taken into account and considerably diminish the default memory allocation done by SynDEx [50].

D. Platform abstraction

This methodology implicitly provides, on one hand an hardware platform abstraction related to the software processing elements, by using already existing IPs. With regards to communications, on the other hand, SynDEx generates a generic code which allows the designer to translate it in whatever language he/she wants, as explained in V-A.

The automatic partitioning is maximum in the case of multi-processing on the SW side (processors) using SynDEx, and on the HW side (FPGAs) using dedicated hardware tools such as GAUT for instance. Only the border between HW and SW has to be manually decided: this is the current limitation of the HS/SW co-design issue.

This abstraction permits ease in terms of portability for new processing units or new platforms. It relies on the use of libraries using the appropriate drivers for each platform.

An IP can be moved from SW to HW if its code is available for both targets, in C and VHDL for example. This allows design adaptation by mean of a simple click:
- migrate a processing element from a processor to a FPGA and vice-versa,
- add a processing element in a data-flow graph whatever is on the processor: a FPGA or an ASIC
- change of platform,
- migrate a processing element from a gate-oriented implementation to a processor-oriented implementation on a processor core inside a FPGA,
- etc.

E. Methodology principles summary

The proposed methodology is based on the SynDEx tool in association with several concepts, such as the component-based approach. This paragraph aims at recalling its advantages and show how it meets the requirements described in part II as for SDR design, and as for the realistic design flow, described in part IV-B.

SynDEx provides an overall application execution time prediction in the context of a multi-processing platform, featuring several possibilities towards heterogeneous design. At least some FPGA design contexts are fully integrated in the methodology, and ASICs are also included in the design. SynDEx has the advantage to propose an optimized and static scheduling, which is a guarantee for hard real-time SDR constraints. Moreover, the methodology allows the automatic code generation for each processing unit of the platform, including IPs encapsulation and communication glue. Since this code is generic (M4 macro-code), it may be translated in whatever language, thus enabling the support of a heterogeneity of processing devices. The re-usability of IPs is intrinsic to the methodology. All this permits to speed up designs at an amazing rate.

The reduced set of non-functional parameters is considered as a positive feature, with regards ultra-fast prototyping. SynDEx is a tool which can generate prototypes based on existing previous work in a couple of days. Students can use it for small projects. Only a few weeks are needed to become an expert. It is affordable to student trainees for a period of several months as soon as a supporting designer is in the lab. Section VI proves the approach efficiency for many design scenarios examples. If a frequent user of this methodology has a hardware platform at his disposal, then he/she gets used to perform the simulation, test and validation of the application on the platform itself.
Please note that, in the future, any other tools providing such extended features, concerning the presented goals compared to SynDEx, would be worth to be considered. Moreover, this methodology could be also expanded at the top of the design flow, towards higher level design tools based on UML. The input graphs from SynDEx could easily be generated from previous graphs defined in a UML environment. A bridge between A3S graphs and SynDEx graphs for instance would be straightforward. SynDEx would then only take the timing non-functional parameter from A3S. The other non-functional parameters (power consumption, memory and surface considerations) taken into account in A3S would not be used in the mapping process itself but would at least have been considered at the very beginning of the design, which is better than nothing. Of course, the long term goal is to really integrate all non-functional parameters in the complete design flow.

VI. DESIGN EXAMPLES (IMPLEMENTATION)

Whatever the methodology is, the necessary SDR development efforts are, on one hand, the coding of an application in order to simulate and validate the correct radio functionality, and on the other hand, the coding of a low level software associated with the hardware components in order to prototype the radio on the platform. The lack of methodology results in the repetition of this process from scratch for each new design. This is the current SDR domain situation.

The methodology suggested in this paper permits to share development efforts for both:
- radio signal processing IPs,
- low level software of hardware architecture.

This is the main cause of the acceleration of the methodology; together with a set of concepts particularly adequate for SDR design and explained earlier. We suggest in the following paragraphs to illustrate the methodology efficiency in various design scenarios. Please note that it will be illustrated that SDR design is tackled in its widest acceptance here. This involves also cross-layer design, as PHY (radio) layer, as well as application (video) layer may be merged with the proposed design approach. This proves also the relevance of the approach for general heterogeneous embedded design outside SDR.

A. Starting from scratch

Our first attempt to develop an SDR is as simple as a broadcast FM receiver development. This is a play activity for a hundred percent software radio implementation, since the antenna is directly connected to the analogue to digital converter. The selected platform is a Pentek P4292 made of four TMS320C6203 DSPs. All this processing power is much greater than necessary, and is only used to evaluate the methodology in a multi-processing context. First of all, the work consists in developing the FM receiver data-flow graph with a component based approach, and the fixed-point C language code for the content of each IP. This represents a time period ranging from a few days to a couple of weeks of work at the most, depending on the previous knowledge of the modulation, the receiver quality (stereo or mono, choice of demodulation algorithm, etc.). This can be done on a first hardware platform architecture made of one PC, which is available by default in SynDEx.

Secondly, low level software for the HW component support are generated for the Pentek platform and gathered in libraries. This concern all the specific primitives, which address synchronisations and threads for the C6203, as well as the communications dedicated to the platform FIFO. This may take several weeks, but less than 2 months and includes M4 learning.

It is not at this stage that the methodology allows to gain anything, since this preliminary work is always necessary. However, as the design steps are very well identified and separated in the methodology, this is already helpful and enables to save time. The work can also be easily separated between two designers, one signal processing designer for the software application (FM) and one for the platform software support.

![Fig. 14. SynDEx FM receiver application graph.](image-url)
(libraries). Fig. 14 gives a FM receiver application graph which enables to generate, with the use of this methodology, the full software radio FM receiver. Bottom Fig. 14 represents the hierarchical view of the top FM Demodulation box.

The FM receiver shown here is a complex one which operates stereo demodulation. Only one DSP is sufficient to run the FM receiver in real-time. Nevertheless the methodology enables to run it on several DSPs with no additional effort.

B. Digital ASIC implication

Under-sampling techniques are used with a 65 MHz ADC sampling frequency. The tuning is managed by a software, which changes the DDC (Digital Down Converter). These considerations may appear or not in the HW platform graph. The digital ASICs present in such architectures, and in the Pentek platform for example, are usually managed by DSPs. DSPs contain the adequate initialization code, necessary to configure ASICS parameters, such as ADC and DDC here, but also DAC (Digital to Analog Converter) and DUC (Digital Up-Converter) in a transmitter context. The DSP may also change these parameters at run-time. ADC and DDC are represented in a unique bloc in Fig. 15, since the only information the DSP needs to know is which media (BIFO_DMA) it has to send data to, and which control information was sent to the 2 ASICs. The connection between the 2 ASICs is described thereafter.

C. A new SW application installed on the previous platform

The Pentek platform, featuring four C6203, is of course supposed to support much more demanding radio applications, such as 3G radio waveforms. We then develop a UMTS FDD baseband chain, implementing DPDCH (data) and DPCCH (control) channels until the intermediate frequency (IF).

After a fixed point functional validation on a PC environment (Visual C++ or Borland), the application is then ported in the embedded multi-DSP platform environment at no cost, since libraries were already existing from FM first experiments. From a practical point of view, this means that once the SW application graph is done in SynDEx and validated on a mono-processor PC architecture graph, the designer just need to take the hardware platform graph of the Pentek platform previously developed for the FM application, and map in one click the UMTS application on the Pentek multi-DSP platform. The methodology automatically generates, thanks to SynDEx and the M4 macro-translator, the main.c executive file of each DSP, including IPs launching and synchronizations means. The designer may even choose how many DSPs are needed to run the application in real-time, thanks to the overall execution time prediction provided by SynDEx. Anyway, he/she can directly check it on the platform itself in real conditions. SynDEx helps indeed to make this prediction while offering the possibility to automatically insert timing measurement between all IPs. Then mono-processing runs are carried out for each potential device target, and measurements provide the execution time figures for each IP. Once these figures have been entered in SynDEx, as well as the communication timings, the tool is able to optimize the multi-processor implementation mapping. It provides a timing prediction and the associated code for each processing unit. The designer may explore different architectural possibilities until he/she is satisfied.

Fig. 16 shows the UMTS transmitter application graph within SynDEx environment. Several hierarchical blocks are present in this graph, as well as repetitions. The hardware architecture then comprises 130 instances and this for a single frame execution. One UMTS frame is made of 15 slots, the content of which has been detailed in the bottom half part of Fig. 16.

Fig. 17 shows the Pentek platform HW architecture with four TI C6203 DSPs and one ADC.

D. Very fast platform extension

We saw in the previous paragraphs that the UMTS application porting on the Pentek platform is done at no cost, since SynDEx libraries already exist from the FM case study. It may also be necessary sometimes to rapidly investigate which little changes could happen. Let us consider the context of the UMTS re-partition processing between a GPP and several DSPs. In the Pentek platform environment, this means adding a TCP link between one of the platform DSP and the host PC processor. SynDEx already has the necessary TCP library on the PC side. So one only needs to develop the DSP side equivalent library, which only consists in encapsulating the
code provided by Pentek in a SynDEx library. This takes one day.

Fig. 18 shows the SynDEx graph of the new platform involving both PC and DSP, which allows now to automatically provide the code and run the UMTS application in a new environment made of one Pentium and four TI C6203 processors.

This gives the opportunity to rapidly evaluate if new design choices are relevant or not, and this at very low cost and in real conditions. This is exactly the opposite of a manual design methodology, where any architecture exploration is very expensive.

E. Very fast application extension

The platform extension regarding the general purpose processing side is experimented in a cross-layer design mixing video and radio processing. We consider the following application already discussed in [39]: a combined UMTS FDD / MPEG-4 implementation.

This cross-layer implementation benefits from two designs made in two different research contexts, but using the same methodology suggested here. On one hand, the SDR design approach we are illustrating in this paper, and on the other hand an image processing perspective to investigate MPEG-4 coding features. These two designs are being developed in 2 completely distinct manners in 2 distinct SynDEx projects. Each of the projects just has to connect to the same TCP socket, on the PC side for the MPEG-4 application, and on the DSP board side for the UMTS radio. Connection is obtained automatically between the two applications, and MPEG-4 data are transmitted through a UMTS link.

F. Rapid prototyping on a new platform

Now the porting context of this complex application gathering UMTS and MPEG-4 applications is considered. This is yet another issue, which needs to be addressed by SDR design. It has been illustrated in [39] in the case of a Pentek platform made of two C6203 DSPs, and a Sundance platform featuring two C6416 DSPs. We suggest referring to [39] for more information about hardware platform description model, and corresponding developed libraries. The schematic of Fig. [19] displays a porting scenario summary.

Another platform used in this work is a desktop computer associated with the VP3-pmC multi-DSP board. The PC consists of an Intel Core 2 Duo CPU at 2.2 GHz. The VP3-pmC is a parallel programmable processing platform dedicated to professional video applications, like MPEG4-AVC/H.264 real-time encoding and MPEG2 to H.264 trans-coding. This platform comprises 5 DSPs TMS320DM642 running at a 720 MHz clock rate, each of them comprising a 32 MBytes SDRAM and a FPGA hardware co-processor, in order to manage the communications between all platform DSPs and the communications with the PCI-Host. All transfers between DSPs are managed by 5 DMA controllers, which are inside the FPGA.

This platform has been previously described with the SynDEx tool in [51]. The UMTS application is automatically prototyped on this platform, and takes advantages of the methodology: no deadlocks, functional checking of the application portability, and automatic multi-processors implementation. Thanks to the small internal memory, the TI cache memory...
is automatically used as presented in [52]. Timings, given in Table III, have proven that increasing the number of DSPs for this application was not accurate, since we obtain only a 1.35 acceleration factor from one to two DSPs, and then no more acceleration whatever the number of DSPs are, and this until there are five of them.

Results for this platform are slower than the compulsory standard real-time figure, which is 10ms. This shows that the application graph has to be redefined, in order to extract more potential parallelism so that the hardware platform can perform more parallel executions.

G. FPGA implication
Another extension of the design space exploration is to combine SW and HW processing, with regard to processors and FPGAs. This could be the solution for the real-time issue encountered in the previous paragraph scenario.

Thanks to the development of VHDL M4 libraries, the suggested methodology enables to generate VHDL as easily as C language code from SynDEx. The sole limitation has been exposed earlier: SynDEx is not appropriate to automatically optimize the scheduling and partitioning of multiple HW IPs, since it based on a sequential processing model between IPs. But all the other methodology features are valid, such as the automatic code generation of communications and synchronization means. And in the particular case of only one IP inside a FPGA, it means that the restriction is null and the FPGA is used as a hardware accelerator.

The UMTS FDD baseband transmitter most demanding IP in terms of processing power is the pulse shaping filter. We propose to map the UMTS receiver in the hardware platform of Fig. 20, in order to speed up the receiver execution. Results of Table IV show the timings obtained after the implementation, without or with an FPGA. An FPGA is used as an accelerating device and permits to respect real-time constraints, since a UMTS FDD frame has to be generated every 10 ms and the combination of a Xilinx Virtex2 FPGA and a TI C62x DSP enables to decrease the frame execution down to 9.9 ms.

H. MPEG-4 Decoder rapid prototyping
An MPEG-4 decoder description has been built for intra pictures in [53], according to the video MPEG-4 texture decoding libraries. The operation granularity of those fits IDCT, VLC and dequantization levels within a block. Description granularity has a significant impact on the final implementation. The MPEG-4 decoder in [53] is extended in [52] to Simple Profile. MPEG-4 natural texture coding tools divide intra or predictive pictures into macroblocks, each made of four 8x8 blocks of luminance channel, and the associated 8x8 blocks of chromatic component. Each MB operation has been optimized for a DSP implementation: interleaving loops, conditional tests leading to a pipeline rupture, no dynamic allocations. Our implementation is an open source one, started with the xvid decoder (http://www.xvid.org), and which received the xvid team agreement to be the first porting over a DSP.

Thanks to the methodology developed here, we would like to emphasize how fast an application can be ported from one platform to another one. In this case, this MPEG-4 decoder has been quickly ported over several platforms presented earlier in this paper: Vitec, Pentek and Sundance. It can decode in real-time sequences up to 2048 Mbps with a VGA resolution at 60 frames per second on the TI C6416 DSP at 1 Ghz. The data-flow application development phase, in atomic operations enabling parallelism expression, requested approximately a full year working time for one person working on a Sundance platform. In comparison, the porting over the 2 other platforms is very fast (less than one day) for a first implementation shot as soon as libraries of Table I in V-A0b are available. This proves the efficiency of the design methodology suggested in this paper. For instance, the development of libraries for the Pentek and Vitec platforms takes around 2 weeks time. Library indeed consists in encapsulating the board support package provided by the platform manufacturer. Note that after the first implementation on a new platform obtained in a few mouse clicks with SynDEx, it takes a few more days to optimize execution time if real-time is not reached at first shot.

<table>
<thead>
<tr>
<th>Target</th>
<th>Sundance</th>
<th>Pentek</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration</td>
<td>1*C64x</td>
<td>1*C62x</td>
</tr>
<tr>
<td>Time/frame</td>
<td>15.9ms</td>
<td>20.2ms</td>
</tr>
<tr>
<td>MFL ratio</td>
<td>60%</td>
<td>84%</td>
</tr>
</tbody>
</table>

**Table III**

**Timings of the UMTS transmitter for a new platform**

<table>
<thead>
<tr>
<th>DSPs</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.3 ms</td>
</tr>
<tr>
<td>2</td>
<td>11.3 ms</td>
</tr>
<tr>
<td>3</td>
<td>11.3 ms</td>
</tr>
<tr>
<td>4</td>
<td>11.3 ms</td>
</tr>
<tr>
<td>5</td>
<td>11.3 ms</td>
</tr>
</tbody>
</table>

**Table IV**

**Timings of the UMTS transmitter without and with FPGA on the Sundance platform**

**Fig. 20. SynDEx Sundance+FPGA hardware platform graph.**
VII. Conclusion

We have discussed and tried to convince the reader of the advantages of the suggested SDR prototyping methodology. This methodology has been derived from years of experience in terms of prototyping, and enables to produce prototypes within ultra short time delays. It can also be used by students as well, thanks to its simplicity.

The main point to be stressed in this methodology is that, the gathering within a unique framework of several concepts and tools, can and do work indeed. Practical limits are defined and clear objectives are targeted. A component based approach is the basis of the methodology, and combined with efficient and realistic automatic transformations, it allows us to propose a heterogeneous design strategy for SDR prototyping, and in a wider perspective, the design of any embedded equipments.

We are convinced that SDR design will not be solved by a brand new design methodology. Moreover, SDR design perspectives are as numerous as radio design perspectives currently are. Other tools may be chosen with regards the rules given in Fig. 4. Moreover, other tools will have to be added to the suggested flow, in order to enhance its current capabilities. In particular, we are investigating higher level design techniques, using UML modelling solutions. Benefits from automatic co-design advances in general will help as well.

A future step of the design solution we suggest, consists in merging the design methodology exposed in this paper with the reconfiguration management that we are investigating in other studies. This will be the SDR design ultimate level we aim to attain.

REFERENCES


Christophe Moy was born in 1972 in La Roche sur Yon, France. He is an INSA engineer (INSA stands for: National Institute of Applied Sciences), Rennes, France, and graduated in 1995. He received his M.Sc. and Ph.D. degrees in Electronics in 1995 and 1999 at the INSA institute. He worked from 1995 to 1999 on spread spectrum and RAKE receivers for the IETR (Institute on Electronics and Telecommunications of Rennes). He then worked 6 years for the Mitsubishi Electric ITE-TCL research lab, where he was focusing on Software Radio systems and concepts, including digital signal processing, HW and SW architecture, co-design methodology and reconfiguration. He also addressed the design of SDR impulse UWB systems. He represented Mitsubishi Electric at the SDR Forum and worked on the French research program A3S (design of SDR through a Model Driven Architecture UML-based methodology), as well as the IST project E+R phase 1. He is now a Professor in SUPELEC. He carries his research in the IETR based SCEE lab, which focuses on Software and Cognitive Radio. He addresses heterogeneous design techniques for SDR, as well as cross-layer optimization topics. He is involved at the European level in the IST Network of Excellence NEWCOM++, and SEC EULER project. He has also been participating to several French ANR project on SDR design, named Idromel and Mopcom.

Mickaël Raulet was born in 1977 in Saint-Brieuc, France. He received his postgraduate certificate in signal, telecommunications, images, and radar sciences from Rennes University in 2002, and his Engineering degree in electronic and computer engineering at the National Institute of Applied Sciences (INSA), Rennes Scientific and Technical University. Then, in 2006, he received a Ph.D. degree in electronic and signal processing, which was delivered by the INSA institute in collaboration with the software radio team of Mitsubishi Electric ITE (Rennes –France). He is currently working at the Institute of Electronics and Telecommunications of Rennes (IETR), as a research engineer focusing in rapid prototyping of video standard compression algorithms on embedded architectures (multi-DSP architecture) and he is also the Image Group project leader of French Media and Network projects, such as Scalim@ges (MPEG-4 SVC main topic) and Mobim@ges (DVB-H main topic). Since 2007, he has been involved in the ISO/IEC JTC1/SC29/WG11 standardization activities (better known as MPEG), such as a Reconfigurable Video Coding Expert.

His interests include video standard compression and telecommunication algorithms and rapid prototyping on multi-DSP architectures from Texas Instruments.
Game Theory in Wireless Communications with an Application to Signal Synchronization

Giacomo Bacci and Marco Luise

Abstract—This paper is concerned with the general issue of game-theoretic techniques applied to the problem of resource allocation in wireless communication networks. Specifically, its first part is devoted to a tutorial explanation of game theory in the context of CDMA wireless networking, whilst the second part focuses on the particular issue of allocating power resources to optimize the receiver performance in terms of spreading code acquisition. The problem of initial signal acquisition is formulated as a noncooperative game in which each transmitter-receiver pair in the network seeks to maximize a specifically chosen utility function. For the problem at hand, the most significant utility function is represented by the ratio of the probability of signal detection to the transmitted energy per bit, and the game each receiver plays consists in setting its own transmit power and detection threshold, under a constraint on the maximum probability of spurious code locks. This formulation of the game captures the tradeoff between obtaining good code acquisition performance and saving as much energy as possible. Using the techniques introduced with the “toy examples” in the first part of the paper, the Nash solution of the proposed game is investigated and found. Closed-form expressions for the optimal transmit power and detection threshold at the Nash equilibrium are derived, and they are compared with simulation results for a decentralized resource control algorithm.

Index Terms—Game theory, Nash equilibrium, resource allocation, power control, multiuser wireless networks, CDMA, synchronization, code acquisition.

I. INTRODUCTION

As is known, game theory is a broad field of applied mathematics that aims at describing and analyzing interactive decision processes. It in fact provides analytical tools to predict the outcome of complex interactions among rational entities, where rationality calls for strict adherence to a strategy based on perceived or measured results [1]. Traditionally, the main areas of its application have been economics, political science, biology, and sociology. But, since the early 1990s, engineering and computer science have been added to the list.

Wireless communications is a suitable scenario for the application of game theory. Since the early days of wireless communications, the importance of radio resource management (RRM) has in fact emerged as a key issue in network design. Cochannel interference, which is due to the shared nature of the wireless medium, represents a major impairment to the performance of wireless communications. The resource competition can be investigated by modeling the network as an economic system, in which any action taken by a user affects the performance of others as well: just the main field of application of game theory. For instance, there is a substantial literature on power control techniques based on noncooperative game theory, mostly focused on data detection for code division multiple access (CDMA) wireless communications networks (e.g., [2]–[26]). The first part of this paper is devoted to a review of game theory and its applications for power allocation, followed by a simple (“toy”) case study of power control for CDMA, that is expedient to explain the main concepts just reviewed (players, utility, Nash equilibrium, Pareto optimality, etc.)

On the other hand, in addition to data detection, every uplink receiver in the base station of a CDMA network must perform the fundamental function of initial signal synchronization to lock onto the terminal’s signature code. The conjecture we try to (dis)prove here is that the optimum decentralized power allocation devised for data detection leads also to a resource allocation strategy that is good for such synchronization function. To the best of our knowledge, similar approaches focusing on synchronization performance have not been investigated in the literature. In particular, we will develop here a game-theoretic analysis to address the problem of optimal resource allocation in a CDMA wireless network so as to improve the performance of code synchronization. Due to the multiple access interference (MAI) caused by the concurrent presence of terminals sharing the same medium, the performance in terms of code synchronization accuracy is affected by the transmission parameters of all users in the network (e.g., transmit power, modulation format, spreading factor). We will introduce and analyze a noncooperative (distributed) game in which the users in the network are allowed to tune their terminals (at both the transmitter and receiver side) to maximize the ratio between the probability of detection to the transmitted energy per bit.

The remainder of the paper is structured as follows. Section II introduces the basics of noncooperative game theory by means of some expedient examples, and sets the basis for this work. The system model is given in Section III, whereas the formulation of the problem and the Nash equilibrium of the noncooperative game are described in Section IV. Section V shows some numerical results for the proposed analysis, and some conclusions and perspectives are outlined in Section VI.

II. GAME-THEORETIC FORMULATION OF CDMA POWER CONTROL

After the seminal contributions to noncooperative game theory given in the 50’s by J. F. Nash (in particular the proof of the existence of a strategic equilibrium for noncooperative
players - the so-called the Nash equilibrium [27, 28]), game theory has been a subject of considerable study, especially in the traditional areas of economics and political science.

More recently, game theory has also been used in telecommunications and wireless communications [29]–[36]. The reason for this blooming of game-theoretic applications lies in the nature of typical interactions between users in a wireless network. The wireless terminals can in fact be modeled as players in a game competing for the network resources (e.g., bandwidth and power), which are typically scarce. Any action taken by a user affects the performance of other users as well. Thus, game theory turns out to be a natural tool for investigating this interplay.

In the next subsections, we will provide some motivating examples for its use in the context of power control for multiaccess wireless networks, also introducing some key definitions and the fundamental analytical tools.

A. Noncooperative games

A (strategic) game consists of three components: a set of players, the strategy set for each player, and a utility (payoff) for each player measuring its level of satisfaction [37]. In its mathematical formulation, the game can be represented as \( G = [\mathcal{K}, \{\mathcal{A}_k\}, \{u_k(a)\}] \), where \( \mathcal{K} = \{1, \ldots, K\} \) is the set of players; \( \mathcal{A}_k \) is the set of actions (strategies) available to player \( k \); and \( u_k(a) \) is the utility (payoff) for player \( k \). The utility depends not only on its own strategy \( a_k \in \mathcal{A}_k \), but also on the actual strategies taken by all of the other players, denoted by \( a_{\setminus k} = (a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_K) \). Hence, \( u_k(a) = u_k(a_k, a_{\setminus k}) \).

Since our focus is on distributed schemes, we concentrate on noncooperative games, where each player \( k \) chooses its strategy \( a^*_k \) to unilaterally maximize its own utility \( u_k(a) \):

\[
\begin{align*}
  a^*_k &= \arg \max_{a_k \in \mathcal{A}_k} u_k(a) \\
  &= \arg \max_{a_k \in \mathcal{A}_k} u_k(a_k, a_{\setminus k}),
\end{align*}
\]

where the latter notation emphasizes that the \( k \)-th player has control over its own strategy \( a_k \) only. In other terms, \( a^*_k \) represents player \( k \)'s best response to the concurrent actions of the other players. For these games, we first need to introduce two fundamental concepts, namely, Nash equilibrium and Pareto optimality [37].

Definition 1: A Nash equilibrium (NE) is a set of strategies \( a^* = (a^*_k, a^*_{\setminus k}) \) such that no user can unilaterally improve its own utility, i.e.,

\[
  u_k(a^*_k, a^*_{\setminus k}) \geq u_k(a_k, a^*_{\setminus k}) \quad \forall a_k \in \mathcal{A}_k, k \in \mathcal{K},
\]

where obviously \( a^*_{\setminus k} = (a^*_1, \ldots, a^*_{k-1}, a^*_{k+1}, \ldots, a^*_K) \). In other words, an NE is a stable outcome of a game in which multiple agents with conflicting interests compete through self-optimization and reach a point where no player has any incentive to unilaterally deviate (whence stability).

Definition 2: A set of strategies \( a = (\tilde{a}_k, \tilde{a}_{\setminus k}) \) is Pareto-optimal if there exists no other set \( a = (a_k, a_{\setminus k}) \) such that \( u_k(a_k, a_{\setminus k}) \geq u_k(\tilde{a}_k, \tilde{a}_{\setminus k}) \) for all \( k \in \mathcal{K} \) and \( u_k(a_k, a_{\setminus k}) > u_k(\tilde{a}_k, \tilde{a}_{\setminus k}) \) for some \( k \in \mathcal{K} \). In other words, when all players settle onto a Pareto-optimal strategy, no player can improve its own utility without reducing the utility of at least another player.

Our focus throughout this work is on pure (i.e., deterministic) strategies. However, players can also opt for mixed (i.e., statistical) strategies. In this case, each player chooses its strategy according to a probability distribution that is known to the other players. Nash proved that a finite noncooperative game always has at least one mixed-strategy NE [27], [28]. This means that a noncooperative game may have no pure-strategy equilibria, one pure-strategy equilibrium, or multiple pure-strategy equilibria. We can also easily show that there could exist more than one Pareto-optimal solution, and that, in general, an NE does not correspond to a Pareto-optimal strategy.

To illustrate the intuitive meaning of these concepts, we consider a trivial example of a static\(^1\) noncooperative game, that we call the near-far effect game. Two wireless terminals (player 1 and player 2) transmit to a certain access point (AP) in a CDMA network. Player 1 is located close to the AP, while player 2 is much farther away, as is depicted in Fig. 1. Hence, \( K = 2 \) and \( \mathcal{K} = \{1, 2\} \). Each user is allowed either to transmit at a certain power level \( p_k \), or to wait \( (p_k = 0) \). This translates into \( \mathcal{A}_k = \{p_k = 0, p\} \). Each terminal achieves a degree of satisfaction which depends on the outcome of the transmission and on the expenditure in terms of the

\(^1\)A game is said to be static if there exists only one time step, which means that the players' strategies are carried out through a single move [37].
energy spent to transmit at power $p_k$. Mathematically, this translates into an adimensional utility $u_k(a) = u_k(p_1, p_2) = t_k - c_k$, where $t_k = 1$ if the transmission is successful and $t_k = 0$ otherwise, and where the cost is $c_k = c \ll 1$ if the player chooses to transmit, and $c_k = 0$ otherwise.

Due to the near-far effect, sketched in Fig. 1, whenever the near player (player 1) chooses to transmit, its transmission is successful irrespective of the action of the far player (player 2). In particular, if $p_1 = p$, player 1 can deliver its information, thus receiving a utility $u_1(p, p_2) = 1 - c$ (irrespective of $p_2$). When on the contrary player 1 is idle, i.e., $p_1 = 0$, its own utility is $u_1(0, p_2) = 0$, irrespective of $p_2$ again. Let us now focus on player 2. Because of the interference caused by player 1, player 2 can only successfully transmit when player 1 is idle ($p_1 = 0$). In this case, $u_2(0, p) = 1 - c$. If both players transmit $p_k = p$, due to the near-far effect, player 2’s transmission fails and $u_2(p, p) = -c$. Similarly to player 1, $u_2(p_1, 0) = 0$ (irrespective of player 1) when player 2 is idle. The near-far effect game above is summarized in the payoff matrix of Fig. 2. Player 1’s actions are identified by the rows and player 2’s by the columns. The pair of numbers in the box represents the utilities $(u_1(p_1, p_2), u_2(p_1, p_2))$ achieved by the players.

To predict the outcome of the near-far effect game, it is fundamental to assume that both players i) are rational, and ii) know each other’s payoff. By inspecting the payoff matrix, it is apparent that player 1’s best strategy is represented by $p_1 = p$ whatever $p_2$ is, since $1 - c < 0$ under the assumption $c \ll 1$. This is known to player 2 as well. Hence, to “limit damage”, he/she rationally chooses to play $p_2 = 0$. As a conclusion, the near-far effect game has only one pure-strategy NE, represented by the strategy $a = (p, 0)$ (the same conclusion follows from Definition 1).

By applying Definition 2, this game can be shown to have two Pareto-optimal solutions, namely $(p, 0)$ and $(0, p)$, corresponding to transmission of player 1 only or player 2 only, respectively. It is true that the (only) pure-strategy NE solution $(p, 0)$ is also Pareto-optimal, but it is also true that i) it is highly unsatisfactory for player 2 since he/she cannot convey any information to the AP, and that ii) the other Pareto-optimal solution can not be attained by a noncooperative game (since it is not a NE). We take this apparent need for fairness as our motivation to introduce power control.

**B. Power control as a static noncooperative game**

Let us provide our near-far effect game with a naive form of power control. Assume now that each terminal is allowed to transmit choosing between two levels of transmit power different from the previous ones, namely, either at a certain amount $p$, or at a reduced level $\mu p$, $0 < \mu < 1$. The power control factor $\mu$ is such that the received power for both players is the same when the far player uses $p$ and the near player uses $\mu p$. Hence, $A_k = p_k = (\mu p, p)$. Similarly to the previous game with no power control, $u_k(p_1, p_2) = t_k - c_k$, where $t_k = 1$ if the transmission for player $k$ is successful, and $t_k = 0$ otherwise, and where $c_k$ is proportional to the consumed energy, i.e., $c_k = c$ if $p_k = p$, and $c_k = \mu c$ if $p_k = \mu p$. As before, due to the near-far effect, player 1 can successfully transmit irrespective of $p_2$, whereas player 2 can correctly reach the receiver only if $p_2 > p_1$. The payoff matrix for this game is shown in Fig. 3. Since $1 - \mu c > 1 - c$, player 1’s best strategy is $p_1 = \mu p$. Consequently, player 2 plays $p_2 = p$. This game has thus one pure-strategy NE, which is also the only Pareto-optimal solution, and player 2 now succeeds to go through when player 1 is idle.

This power control technique seems to compensate for the near-far effect, since both players are now able to transmit, although player 1 still tends to dominate player 2. However, this scenario does not actually model real data networks. The main inaccuracy lies in the over-simplified “go/no-go” utility function that does not take into account the actual signal-to-interference-plus-noise ratio (SINR) achieved at the receiver. In a data network, higher SINRs lead to a larger amount of transmitted information. This implies that the utility for a data terminal is a continuous function of its achieved SINR.

To account for this different point of view, the term $t_k$ should be a function of the amount of information that is actually delivered to the receiver. Focusing on player 1, if $p_1 = p$, the (normalized) amount of information (we may call it the throughput) is equal to $t_1 = t \gg 0$. If player 1 uses a lower power $p_1 = \mu p$, then $t_1 = \mu t$, with $\mu < \lambda \leq 1$.\(^3\) Considering player 2, $t_2 = 0$ if $p_2 \leq p_1$, and $t_2 = \lambda t$ if $p_2 > p_1$, since the received power for player 2 is equal to that of player 1 with $p_1 = \mu p$.

The payoff matrix for this more realistic game is now shown in Fig. 4. As before, player 1’s best strategy is represented by $p_1 = p$ whatever $p_2$ is, since $t - c > \lambda t - \mu c$ under the assumption $t \gg c$. As a consequence, player 2 rationally chooses to play $p_2 = \mu p$. The pure-strategy NE is represented by the strategy $(p, \mu p)$, whereas the Pareto-optimal solutions are $(p, \mu p)$ and $(\mu p, p)$. We appear to be back to the original situation we had without power control. However, we can make this situation considerably fairer by introducing a new

\(^2\)This hypothesis involves the concept of complete information [37].

\(^3\)Note that $\mu < \lambda$ in all practical scenarios, since the performance in terms of correct detection does not show a linear dependence on the transmit power.
application of a noncooperative game wherein the users are allowed to choose their transmit powers according to a utility-maximization criterion, where the utility is defined as the ratio of throughput to transmit power. Although the focus of the game described in the following is on data detection, the energy-efficient paradigm also applies to the game-theoretic formulation derived in the next sections and tailored to the problem of code synchronization for multiaccess wireless data networks.

Let $G = [\mathcal{K}, \{P_k\}, \{u_k(p)\}]$ be the dynamic noncooperative power control game, where $\mathcal{K} = \{1, \ldots, K\}$ is the index set for the terminal users of the multiuser wireless network; $P_k = [0, P_k]$ is the strategy set; and $u_k(p)$ is the payoff function for user $k$. Note that, unlike previous examples, here the allowed transmit powers can take all the values in the continuous interval $[0, P_k]$.

In terms of the utility function, several different definitions were investigated in the literature, depending on the goal to be achieved [3], [10], [13]–[15]. When energy efficiency is the main concern, as it is here, there exists a tradeoff between obtaining high SINR levels and consuming low energy. A good way to assess this tradeoff is through the number of bits that can be transmitted per joule of energy consumed [2], [4]. This can be quantified by defining the utility function of the $k$th user to be the ratio of its throughput $T_k$ to its transmit power $p_k$. i.e.,

$$u_k(p) = u_k(p_k, p_k) = \frac{T_k}{p_k},$$

where $p = (p_1, \ldots, p_K) = (p_k, p_k)$ represents the vector of transmit powers, with $K$ denoting the number of users, and $p_k$ representing the vector of elements of $p$ other than the $k$th element. The throughput $T_k$ (sometimes referred to as goodput) can be expressed as

$$T_k = R_k f(\gamma_k),$$

where $R_k$ and $\gamma_k$ are the transmission rate and the received SINR for the $k$th user, respectively. The function $f(\gamma)$ is known as the efficiency function, which expresses the packet success rate (PSR), i.e., the probability that a packet is received without an error, as a function of SINR. We assume that a packet is retransmitted if it has one or more bit errors, i.e., no forward error correction (FEC) techniques are considered. Of course, $f(\gamma_k)$ depends on the details of the physical layer, including modulation, coding, and packet size. However, in most practical cases, the efficiency function shows some common features: $f(\gamma)$ is increasing, $S$-shaped (sigmoidal), continuously differentiable, with $f(0) = 0$, $f(+\infty) = 1$, and $f'(0) = df(\gamma_k)/d\gamma_k|_{\gamma_k=0} = 0$ [38]. Combining (3) and (4), we can write

$$u_k(p) = R_k f(\gamma_k),$$

This utility function, which has units of bits/joule, represents the total number of correct data bits that are delivered to the destination, and captures the tradeoff between data rate and battery life. Fig. 5 shows a typical shape of the utility function in (5) as a function of transmit power of user $k$ keeping all of the other users’ transmit power constant. The
utility function in (5) can also be used for coded systems by modifying the efficiency function \( f(\gamma_k) \) to represent the PSR for coded systems and also scaling the transmission rate to count only the information bits in a packet.

Recalling (1), and considering \( p_k = [0, \tau_l] \), player \( k \)'s best response \( r_k(p_{\cdot,k}) \) to a given interference vector \( p_{\cdot,k} \) is [5]

\[
r_k(p_{\cdot,k}) = \min(\overline{p}_k, p_k^*),
\]

where

\[
p_k^* = \arg \max_{p_k \in \mathbb{R}^+} u_k(p_k, p_{\cdot,k}) = \arg \max_{p_k \in \mathbb{R}^+} R_k \frac{f(\gamma_k)}{p_k}
\]

is the unconstrained maximizer of the utility in (5). Fig. 5 illustrates the quantities \( p_k^* \) and \( r_k(p_{\cdot,k}) \).

Clearly, if this value is not feasible, i.e., if \( p_k^* > \overline{p}_k \), then the best action the user \( k \) can choose is transmitting at the maximum value \( \overline{p}_k \).

We can show [38] that in a CDMA network over a flat-fading wireless channel, \( u_k^* \) occurs when the terminal \( k \), using the transmit power \( p_k^* \), achieves the optimal (target) SINR \( \gamma_k^* \), which is a function of its own transmit parameters only. As a consequence, the target SINR \( \gamma_k^* \) can be computed by each player \( k \) before the game starts. Using the analysis developed in [5], (7) can be used to provide an iterative algorithm, which mimics a dynamic game to guarantee a good tradeoff between fairness of the network and efficiency of the resource allocation scheme. At the \((n + 1)\)-th step, user \( k \) updates its (optimal) transmit power \( p_k^*(n + 1) \) according to

\[
p_k^*(n + 1) = p_k^*(n) \cdot \frac{\gamma_k^*}{\gamma_k(n)}.
\]

where \( p_k^*(n) \) is the transmit power at the \(n\)-th step, and \( \gamma_k(n) \) is the SINR experienced by user \( k \) at the \(n\)-th step, that can be fed back by the AP using a return channel. Note that the update process (8) recalls the SINR-balancing criterion derived in the milestone works in the field of distributed power control [39]–[42]. Using the analysis presented in [5], we can show that the Nash equilibrium (7) exists and is unique, although it is not Pareto efficient. Methods to move the Nash equilibrium closer to the Pareto-optimal solution make use of pricing techniques, and can be found in [4], [5]. This algorithm can be extended to the case of frequency-selective scenarios using the analysis presented in [11], [12].

### III. Game-theoretic Power Control for Signal Acquisition

The examples described in the previous section are primarily focused on the issue of data detection for CDMA wireless networks. In the following, we will present a game-theoretic resource allocation scheme based on energy efficiency, but specifically suited to the problem of code synchronization. In this context, the maximization of the achieved throughput per energy consumed is replaced by the maximization of the probability of code acquisition per energy consumed.

In the uplink of a multiaccess CDMA infrastructure network, we consider code acquisition on a pilot channel (i.e., either with no data modulation or modulated with known data), and we assume for simplicity the presence of \( K \) equi-format users with binary signaling and a common spreading factor \( M \) (i.e., the transmission bit rate \( R_b = 1/T_b \) is common among all users, where \( T_b = MT_c \) is the bit time and \( T_c \) is the chip time). The transmission takes place over a frequency-flat and slow-fading additive white Gaussian noise (AWGN) channel. Using baseband-equivalent representation, the signal transmitted by each user \( l \) can be expressed as

\[
s_l(t) = \sqrt{2p_l} \sum_{n=-\infty}^{\infty} g_l(t - nT_b)
\]

where \( p_l \) is the \(l\)-th user’s transmit power, and where

\[
g_l(t) = \sum_{m=0}^{M-1} c_m^{(l)} a(t - mT_c)
\]

is the \(l\)-th user’s signature (bandlimited) waveform. In (9), \( c_l = \{c_m^{(l)}\}_{m=0}^{M-1} \) denotes the spreading code for user \( l \), which is assumed to be random, with \( c_m^{(l)} \in \{\pm 1\} \), and

\[
E\{e_m^{(l)} e_{m+\ell}^{(l)}\} = \begin{cases} 1, & \ell = 0, \\ 0, & \ell \neq 0 \end{cases}
\]

\[
E\{e_m^{(l)} e_{m+\ell}^{(l)}\} = 0, & \forall j \neq l, \forall \ell,
\]

where \( E[\cdot] \) denotes statistical expectation. Also, \( a(t) \) is a square root raised cosine (SRRC) pulse with energy \( T_c \) (the chip shaping pulse). No data modulation is present.

Assuming perfect carrier frequency synchronization, the received signal is

\[
r(t) = \sum_{l=1}^{K} h_l e^{j\theta_l} s_l(t - \tau_l) + \eta(t),
\]

where \( h_l, \theta_l, \) and \( \tau_l \) are the attenuation, the phase offset, and the delay, respectively, experienced by the \(l\)-th user’s
Fig. 6. Serial search architecture for user $k$’s code shift $\Delta_k$.

signal when propagating through the wireless channel; and $\eta(t)$ represents the zero-mean complex-valued AWGN with two-sided power spectral density (PSD) $2N_0$. To simplify our problem, we concentrate on a chip-synchronous scenario, i.e., the unknown signal delay to be estimated is an integer multiple of the chip interval $T_c$; $\tau_l = \Delta_l T_c$ for every $l$, where the code shift $\Delta_l$ introduced by the channel is uniformly distributed in $\{0, 1, \cdots, M-1\}$. After chip-matched filtering and sampling, the received signal at the uplink receiver can be represented as

$$x[m] = r(t) \otimes \frac{1}{T_c} \alpha(-t) \bigg|_{t=mT_c} = \sum_{l=1}^{K} h_l e^{j\theta_l} \sqrt{2P_c} m^{(l)} + \nu[m],$$

where $\nu[m] = \nu_f[m] + j\nu_Q[m]$ is Gaussian-distributed, with independent components $\nu_f[m], \nu_Q[m] \sim N(0, \sigma^2)$, and $\sigma^2 = N_0/T_c$ is the noise power of each component. This model can be easily extended to totally distributed systems, such as ad-hoc networks, provided that all channel components $h_l, \theta_l, \tau_l$ are replaced by their counterparts $h_{lk}, \theta_{lk}, \tau_{lk}$, that account for the $l$-th transmitter / $k$-th receiver pair.

Coming back to the infrastructure configuration, in order for the base station to lock the spreading codes of all $K$ users in the network, the receiver is equipped with $K$ detectors to search for the correct code shifts $\Delta_k$ for all $k \in \{1, \cdots, K\}$. The simplest technique that can be employed is the well-known **serial search** sketched in Fig. 6 for user $k$, that is applied since the early days of CDMA [43]–[45]. The sufficient statistics to test code alignment is obtained after despreading and weighing the received samples $x[m]$ as follows:

$$v_{k}[m; \hat{\Delta}_k] = \frac{1}{h_k \sqrt{2p_k}} e^{(k)}_{m+\hat{\Delta}_k}$$

where $\hat{\Delta}_k$ is the tentative code shift of the locally generated sequence $c_k$. Note that the receiver for user $k$ must estimate the $k$’s transmit power $p_k$ and channel attenuation $h_k$. For the sake of analysis, we suppose perfect estimation of both values.

At the output of the despreader, we have

$$y_{k}[m] | \Delta_k = \sum_{l=1}^{K} \frac{h_l e^{j\theta_l}}{h_k \sqrt{2p_k}} m^{(l)} + \nu[m] \frac{|c_{m+\Delta_k}|^2}{h_k \sqrt{2p_k}}$$

and, after accumulation over a code length $M$, we get

$$z_{k}[n] | \hat{\Delta}_k = \sum_{m=0}^{M-1} y_{k}[m] | m = \mu_k e^{j\theta_k} + \zeta^\text{(MAI)}_{k}[n] + \zeta^\text{(AWGN)}_{k}[n],$$

where

$$\mu_k = \begin{cases} 1, & \hat{\Delta}_k = \Delta_k, \\ 0, & \hat{\Delta}_k \neq \Delta_k. \end{cases}$$

and where $\zeta^\text{(MAI)}_{k}[n]$ is the term arising from multiple access interference (MAI) inherent to (asynchronous) CDMA. The last term is due to AWGN and is clearly given by

$$\zeta^\text{(AWGN)}_{k}[n] = \zeta^\text{(AWGN)}_{I,k}[n] + j\zeta^\text{(AWGN)}_{Q,k}[n],$$

where

$$\zeta^\text{(AWGN)}_{I,k}[n], \zeta^\text{(AWGN)}_{Q,k}[n] \sim \mathcal{N}(0, \frac{N_0}{h_k^2 \cdot 2p_k \cdot M}).$$

By virtue of the central-limit theorem, the term due to the MAI can be approximated to a Gaussian random variable:

$$\zeta^\text{(MAI)}_{k}[n] = \zeta^\text{(MAI)}_{I,k}[n] + j\zeta^\text{(MAI)}_{Q,k}[n],$$

with

$$\zeta^\text{(MAI)}_{I,k}[n], \zeta^\text{(MAI)}_{Q,k}[n] \sim \mathcal{N}(0, \frac{\sum_{l \neq k} h_l^2 \cdot 2p_l / 2}{h_k^2 \cdot 2p_k \cdot M}).$$

We can thus statistically characterize $z_{k}[n] | \hat{\Delta}_k$ in a more compact form as follows:

$$z_{k}[n] | \hat{\Delta}_k = z_{I,k}[n] | \hat{\Delta}_k + j z_{Q,k}[n] | \hat{\Delta}_k,$$

$$z_{I,k}[n] | \hat{\Delta}_k \sim \mathcal{N}(\mu_k \cdot \cos \theta_k, \frac{1}{2\gamma_k}),$$

$$z_{Q,k}[n] | \hat{\Delta}_k \sim \mathcal{N}(\mu_k \cdot \sin \theta_k, \frac{1}{2\gamma_k}),$$

where

$$\gamma_k = \frac{E_R^{(e)}}{I_{0,k} + N_0} = \frac{M \cdot h_k^2 p_k}{\sum_{l \neq k} h_l^2 p_l + \sigma^2}$$

is the SINR of user $k$, defined as the ratio between the energy per bit of user $k$ collected at its receiver $E_R^{(e)}$ and the received PSDs due to MAI $I_{0,k}$ and to AWGN $N_0$, respectively.

We are now to detail the **synchronization strategy** shown in Fig. 6. Broadly speaking, this block combines $z_{I,k}[n] | \hat{\Delta}_k$...
and \( z_{Q,k}[n] \) \( \Delta_k \) to derive a final real-valued statistics \( w_k[n] \) \( \Delta_k; \rho_k \) to test acquisition. This quantity is a function of both the tentative code shift \( \Delta_k \) (through the mean value of \( z_k[n] \)), and of the synchronization strategy that we choose, denoted by \( \rho_k \), since the way the components of \( z_k[n] \) are combined impacts on its probability density function (pdf). More details on practical synchronization strategies (e.g., coherent or non-coherent) will be provided in Section V.

To decide whether the \( k \)-th receiver is in-sync or out-of-sync, the output \( w_k[n] \) \( \Delta_k; \rho_k \) is compared with a detection threshold \( \lambda_k[n] \in [0,1] \). For convenience of notation, we will drop the dependence of quantities such as decision statistics and threshold on the symbol index \( n \) from now on. In case the test fails, i.e., if \( w_k[\Delta_k; \rho_k < \lambda_k] \), then a new tentative code shift \( \Delta_k \) is selected for the PN sequence generation (14). If the synchronization test is passed, i.e., if \( w_k[\Delta_k; \rho_k > \lambda_k] \), then the receiver assumes that the tentative delay \( \Delta_k \) of the locally generated PN sequence is the correct delay \( \Delta_k \) and proceeds to verification mode [46], [47]. This process aims at avoiding false code locks, which are extremely detrimental for the receiver in terms of increased time for correct synchronization and subsequent data detection. However robust, verification is expensive in terms of time and processing resources. Hence, a key performance indicator of the acquisition strategy is given by the probability of false alarm, defined as

\[ P_{FA}(\gamma_k, \lambda_k; \rho_k) = \Pr\{ w_k > \lambda_k | \Delta_k \neq \Delta_k; \rho_k \} = \int_{\lambda_k}^{+\infty} f_W(\Delta_k = \Delta_k; \rho_k) (w) dw, \quad (23) \]

to be kept as low as possible. The twin performance parameter that fully characterizes the sync procedure is the probability of detection, defined as

\[ P_D(\gamma_k, \lambda_k; \rho_k) = \Pr\{ w_k > \lambda_k | \Delta_k = \Delta_k; \rho_k \} = \int_{\lambda_k}^{+\infty} f_W(\Delta_k = \Delta_k; \rho_k) (w) dw, \quad (24) \]

which on the contrary should be as high as possible, since a missed detection implies a much longer acquisition time and thus a significant delay in the receiving chain. In our definitions above, we expressed \( P_{FA} \) and \( P_D \) as explicit functions not only of the tentative code shift and of the adopted strategy, but also of the SINR \( \gamma_k \) through the variance of \( z_k \).

From (23)-(24) we can extract general requirements for synchronization strategies to be amenable to our game-theoretic analysis. The first requirement is in terms of the behavior of \( P_{FA} \) and \( P_D \) versus the threshold \( \lambda_k \in [0,1] \) for a fixed SINR \( \gamma_k \), for all \( k \in \{1, \ldots, K\} \): both \( P_{FA} \) and \( P_D \) must decrease as \( \lambda_k \) increases. This is a reasonable assumption, since, for a fixed SINR, increasing \( \lambda_k \) means increasing the lower interval of the integrals (23)-(24), thus reducing the interval of integration. The second requirement is in terms of the behavior of \( P_{FA} \) versus the SINR \( \gamma_k \) for a fixed threshold \( \lambda_k \), for all \( k \in \{1, \ldots, K\} \): \( P_{FA} \) must decrease as \( \gamma_k \) increases, with \( \lim_{\gamma_k \to +\infty} P_{FA} = 0 \). The third requirement is in terms of the behavior of \( P_D \) versus the SINR \( \gamma_k \) for a fixed threshold \( \lambda_k \), for all \( k \in \{1, \ldots, K\} \): \( P_D \) must increase as \( \gamma_k \) increases, with \( \lim_{\gamma_k \to +\infty} P_D = 1 \). Again, the last two requirements appear to be reasonable, since increasing \( \gamma_k \) tightens the pdf around its mean value, which is 0 in case of wrong code shift and 1 in case of correct code shift.

Under these assumptions (that are invariably satisfied by practical synchronization strategies), the performance of the considered system model, measured in terms of probabilities of detection and false alarm, increases as the SINR increases. As a consequence, there exists a tradeoff between good synchronization performance on one side, and low energy consumption on the other. This is analogous to numerous approaches in the field of resource allocation for data detection, as those described in Section II-C, in which the performance index is given by the effective throughput at the receiver.

IV. FORMULATION OF THE CODE-SYNC GAME

The tradeoff between good synchronization performance and energy saving, is captured if we define a utility function as the ratio between the probability of detection \( P_D \) and the transmitted energy per bit (or per acquisition): \( E^{(t)}_k = p_k \cdot T_b \)
as follows:

\[ u_k(p, \lambda_k) = u_k(p_k, \lambda_k), \lambda_k = \frac{P_D(\gamma_k, \lambda_k; p_k)}{p_k \cdot T_b}, \quad (25) \]

where \( p_k = p \setminus \lambda_k = (p_1, \ldots, p_{k-1}, p_{k+1}, \ldots, p_K) \). Each \( k \)-th transmitter-receiver pair can set its own transmit power \( p_k \) (at the transmitter side) and its detection threshold \( \lambda_k \) (at the receiver side). On the contrary, the synchronization strategy \( \rho_k \) is assumed to be given, since it depends on the availability of some a-priori information on the channel experienced by transmitter \( k \), such as the residual phase offset \( \theta_k \), which may or may not be available at the \( k \)-th receiver.

Although \( E^{(t)}_k \) is a function of \( p_k \) only, \( P_D \) depends on the achieved SINR at the receiver in addition to the threshold. Using (22), we can easily verify that \( \gamma_k \) depends not only on the value of \( p_k \), but also on all other users’ transmit powers \( p_{\setminus k} \). As a consequence, maximizing \( u_k(p, \lambda_k) \) is not a unilateral optimization, but a multidimensional problem. In this context, we can formulate a noncooperative game with complete information [1] in which every transmitter-receiver pair seeks to maximize its own utility by choosing an optimum configuration in terms of (transmit power, threshold).

In addition to this, we have also to place a constraint on the (maximum) probability of false alarm \( P_{FA,k} \) to limit the occurrence rate of spurious detections. In some sense, \( P_{FA,k} \) represents user \( k \)’s quality-of-service (QoS) requirement, that depends on the desired accuracy to be achieved. Let \( G = [K, \{A_k\}, \{w_k\}] \) be such game, in which \( K = \{1, \ldots, K\} \) is the index set for the transmitter-receiver pairs; \( A_k = P_k \times A_k \) is the strategy set for user \( k \in K \), where \( P_k \) is the transmit power set, and \( A_k \) is the set of threshold values; and \( w_k \) is the payoff function for the pair \( k \). The power strategy set is \( P_k = [\underline{p}_k, \overline{p}_k] \), with \( \underline{p}_k \) and \( \overline{p}_k \) denoting the minimum and maximum power constraints, respectively. The threshold strategy set is \( A_k = [0,1] \) for all receivers \( k \in K \). Formally,
\[ \mathcal{G} \] can be expressed as
\[
(p_k^*, \lambda_k^*) = \arg \max_{p_k \in P_k, \lambda_k \in A_k} u_k(p, \lambda_k)
\]
subject to
\[
P_{FA}(\gamma_k, \lambda_k; p_k) \leq \mathcal{T}_{FA,k} \quad (26)
\]
where \( u_k(p, \lambda_k) \) is defined as in (25). The optimality criterion (26) can be reformulated more conveniently by inverting (22):
\[
p_k^* = \frac{\sum_{k \neq k} h_k^2 p_k + \sigma^2}{M \cdot h_k^2}, \quad \gamma_k = \frac{\lambda_k}{\xi_k^k},
\]
where \( \xi_k \) is independent of \( p_k \) and \( \lambda_k \). Hence, (26) equals
\[
(\gamma_k^*, \lambda_k^*) = \arg \max_{\gamma_k \in [0, \xi_k \mathcal{T}_{FA,k}], \lambda_k \in [0, 1]} \frac{P_D(\gamma_k, \lambda_k; p_k)}{\gamma_k}
\]
subject to \( P_{FA}(\gamma_k, \lambda_k; p_k) \leq \mathcal{T}_{FA,k} \quad (28) \)
with \( \gamma_k^* = \xi_k p_k^* \).

Now, adding the constraint \( P_{FA}(\gamma_k, \lambda_k; p_k) \leq \mathcal{T}_{FA,k} \) actually means identifying that subset \( \mathcal{A}_k \subset [0, \xi_k \mathcal{T}_{FA,k}] \times [0, 1] \) that provides \( P_{FA}(\gamma_k, \lambda_k; p_k) \leq \mathcal{T}_{FA,k} \). Since \( P_{FA} \) is assumed to decrease as both \( \lambda_k \) and \( \gamma_k \) increase, we can set \( \lambda_k = 1 \) to identify the minimum SINR \( \gamma_k \) that provides \( P_{FA}(\gamma_k, \lambda_k = 1; p_k) = \mathcal{T}_{FA,k} \). For the sake of notation, we define a function \( \gamma(\cdot) \), depending on both the QoS constraint and the synchronization strategy, such that
\[
\gamma_k^* = \gamma(\mathcal{T}_{FA,k}; p_k).
\]

It is easy to verify that, for any \( \lambda_k \in [0, 1] \), a necessary condition for the QoS constraint to be verified is \( \gamma_k \in \left[ \gamma_k^*, \xi_k \mathcal{T}_{FA,k} \right] \).

Given a fixed \( \gamma_k \in \left[ \gamma_k^*, \xi_k \mathcal{T}_{FA,k} \right] \), a subset \( \mathcal{A}_k(\gamma_k) \subset [0, 1] \) that fulfills \( P_{FA}(\gamma_k, \lambda_k; p_k) \leq \mathcal{T}_{FA,k} \) can also be obtained using similar arguments. Since \( P_{FA} \) is a decreasing function of \( \lambda_k, \Delta_k(\gamma_k) \) is such that \( P_{FA}(\gamma_k, \lambda_k = \Delta_k(\gamma_k); p_k) = \mathcal{T}_{FA,k} \).

Similarly, we define a function \( \Delta(\cdot) \) such that
\[
\Delta_k(\gamma_k) = \Delta(\gamma_k, \gamma_k^*; p_k) \quad (30)
\]
that accounts for the QoS constraint and the receiver type. Based on the customary behavior of \( P_{FA}, \Delta_k(\gamma_k) \) increases as \( \gamma_k \) increases, while it decreases as \( \gamma_k \) increases. The functions \( \Delta(\cdot) \) and \( \Delta(\cdot) \) will be explicitly characterized in Section V.

Note that, due to the decreasing monotonic behavior of \( P_{FA} \) with respect to both \( \lambda_k \) and \( \gamma_k \), both \( \Delta(\cdot) \) and \( \Delta(\cdot) \) are bijective functions.

Finally, our game (28) can be cast into the following compact form:
\[
(\gamma_k^*, \lambda_k^*) = \arg \max_{\gamma_k \in [\gamma_k^*, \xi_k \mathcal{T}_{FA,k}], \lambda_k \in [0, 1]} \frac{P_D(\gamma_k, \lambda_k; p_k)}{\gamma_k} \quad (31)
\]
From now on, we will replace \( P_D(\gamma_k, \Delta_k(\gamma_k); p_k) \) with \( P_D(\gamma_k; \gamma_k^*, p_k) \) and we explicitly notice that our formulation includes: i) the energy-efficient tradeoff in terms of good synchronization performance versus energy consumption; ii) the synchronization strategy chosen; and iii) the QoS constraint in terms of probability of false alarm.

It is time now to proceed to solve our resource allocation problem (31) using the analytical tools of game theory. In particular, we search for the Nash equilibria of the game \( \mathcal{G} \), i.e., those pairs \( (p_k^*, \lambda_k^*) \) such that, for all \( k \in K \),
\[
 u_k \left( (p_k^*, \lambda_k^*) \right) \geq u_k \left( (p_k, \lambda_k) \right)
\]
for all transmit powers \( p_k \in P_k \) and for all thresholds \( \lambda_k \in A_k \), and such that \( P_{FA}(\gamma_k, \lambda_k; p_k) \leq \mathcal{T}_{FA,k} \). We will formulate our main results in terms of theorems.

**Theorem 1:** The game \( \mathcal{G} \) admits (at least) one NE if the following two conditions are met:
\[
P_D^0 \left( \tilde{\gamma}_k; \tilde{\gamma}_k, p_k \right) = \frac{d}{d\tilde{\gamma}_k} P_D(\gamma_k; \gamma_k, p_k) < 0 \quad (33a)
\]
\[
\Phi = \sum_{k=1}^K \varphi_k < 1 \quad (33b)
\]
where
\[
\varphi_k = \left( \frac{M}{\gamma_k} + 1 \right)^{-1} > 0, \quad (34)
\]
and where
\[
\gamma_k^* = \begin{cases} 2\gamma_k, & \text{if } \tilde{\gamma}_k < \frac{2\gamma_k}{\lambda_k^*}, \\ \tilde{\gamma}_k, & \text{otherwise} \end{cases}
\]
is the SINR that maximizes \( u_k(p, \lambda_k) \), with \( \tilde{\gamma}_k > \gamma_k^* \) satisfying the relation \( \tilde{\gamma}_k = h \left( \tilde{\gamma}_k; \gamma_k^*; p_k \right) \), and \( h \left( \gamma_k^*; \gamma_k^*; p_k \right) = P_D(\gamma_k^*; \gamma_k^*; p_k) / \left( dP_D(\gamma_k; \gamma_k^*, p_k) / d\gamma_k \right) \).

The proof, omitted for the sake of brevity, is made in two steps, showing that i) \( u_k(p, \lambda_k) \) is continuous and quasi-concave in \( (p_k, \lambda_k) \in \mathcal{A}_k \subset \left[ \gamma_k^*, \xi_k \mathcal{T}_{FA,k} \right] \times [\Delta_k(\gamma_k), 1] \subset \mathcal{A}_k \); and ii) \( \mathcal{A}_k \) is a nonempty, convex, and compact subset of some Euclidean space. A typical \( \mathcal{A}_k \) is given by the shadowed area, including its contour, of Fig. 7. This condition is sufficient to ensure the quasi-concavity of \( u_k(p, \lambda_k) \).

As can be easily verified, (33a) represents a rather loose constraint to the shape of \( P_D \) as a function of \( p_k \). On the contrary, the requirement (33b) is a necessary and sufficient condition for \( \mathcal{A}_k \) to be nonempty, and implies some form of admission control, to be performed before the game starts to
allow all players to achieve their target SINRs $\gamma_k^*$. In the case that $\Phi \geq 1$, there will be a subset $K'$ of players that will achieve $\gamma_k^*$, whilst the remaining $K \setminus K'$ will not. The quantity $\varphi_k$ can be thought of as the size that player $k$ (i.e., the $k$-th transmitter-receiver pair) occupies in the space of the available resources (the bidimensional space $A_k = P_k \times A_k$). We can show that $\gamma_k^*$ (and thus $p_k^*$) increases as $\varphi_k$ increases. Hence, the larger $\varphi_k$, the larger amount of such resources “consumed” [9], [48].

To derive other properties of the NE, including its uniqueness, we consider it from another point of view, focusing on the transmit power $p_k$. The power level chosen by a rational self-optimizing player is its best response to the powers $p_k$ chosen by the other players (any threshold, including receiver $k$’s $\lambda_k$, does not affect this choice, due to the monotonic decreasing behavior of $P_D$ with $\lambda_k$). Formally, player $k$’s best response is the map that assigns each $p_k \in P_k$ the set

$$ r_k (p) = \{ p_k \in P_k : u_k \left( \left( p_k, p_{-k} \right), \lambda_k \right) \geq u_k \left( \left( p_k, p_{-k} \right), \lambda_k \right) \} \quad \text{for all } p_k \in P_k, \quad (36) $$

where $P_k$ is the strategy space of all users excluding user $k$. With the notion of a player’s best response, the transmit power at the NE can be restated in a compact form: $p^*$ is the vector of transmit powers at the NE of the game $\mathcal{G}$ for all $k \in K$.

Theorem 2: The game $\mathcal{G}$ has a unique NE, achieved when

$$ \lambda_k^* = \Delta_k \left( \gamma_k^* \right) \quad (37) $$

$$ p_k^* = \frac{\gamma_k^*}{\xi_k} - \frac{\varphi_k}{h_k} - \frac{\sigma^2}{1 - \Phi} \quad (38) $$

where $\gamma_k^*$ is defined as in (35). The proof, provided for brevity again, can be obtained showing that the correspondence $r^*(p^*) = \left( r_1 \left( p_{1}^* \right), \cdots, r_K \left( p_{K}^* \right) \right)$ is a standard function [42].

It is interesting to note that the expression (38) for the level of transmit power at the NE is also applicable to the case of the data-detection-oriented power control scheme illustrated in Section II-C. Using the analysis reported in [7], we can show that $p_k^*$ is a function of $\varphi_k$ and $\Phi$, provided that the definition (34) makes use of the target SINR $\gamma_k$ in accordance with the utility function (5). So the criterion to perform optimal resource allocation for code synchronization is the same as the one for data detection, but the target SINRs and the resulting power levels may turn out to be different in the two cases, according to the different requirements coming either from data detection or from synchronization. The difference in terms of power levels at the NE between the two cases derives from the fact that we must replace $\gamma_k^*$ that follows from (35) with the SINR that maximizes (5). In the practice, during the initial phase of code synchronization we use $\gamma_k^*$ in accordance with (35), and we jointly set $\lambda_k^*$ and $p_k^*$ following (37)-(38) to maximize the probability of correct code acquisition per transmitted energy consumed, without caring of the achieved throughput at the receiver. After acquisition is over, we will use the target SINR that maximizes (5) to maximize the goodput at the receiver per energy consumed at the transmitter, neglecting the (now irrelevant) performance of the receiver during acquisition.

Similarly to the case of the energy-efficient power control described in Section II-C, the NE for code synchronization is not Pareto-optimal. We could use pricing techniques similar to that employed for data-detection-based schemes, to obtain Nash solutions which are superior in the sense of Pareto-optimality.

V. Simulation results

We provide in this section some numerical results to support and integrate the theoretical analysis presented in Section IV. Our iterative distributed algorithm (not reported in detail for brevity) is a straightforward implementation of the best response outlined above, and it is similar to the update criterion (8) reported in Section II-C for the case of data detection. Note that, using the theoretical formulation of [42], we can easily show that the decentralized best-response algorithm always converge to the NE. The only information that must be provided by the access point, as it is not locally available at the transmitter, is the current received SINR level, which can be fed back by the receiver with a very modest data rate requirement on the signaling channel. An issue still open in the current formulation is the way the receiver measures the SINR level when the code synchronization is not yet achieved. The following simulations are thus conducted assuming a genie-aided framework that allows each transmitter to know its received SINR at the concentration point. The same considerations applies to the case of coherent detectors, since it is very unlikely that a good carrier phase estimate (required by coherent detection) can be obtained prior to achieving code lock and performing signal despeading.

Table I specifies the details of our analysis both for an ideal coherent synchronizer, in which $\theta_k$ is perfectly known at the receiver, and for a non-coherent synchronizer, in which $\theta_k$ is unknown (in the proposed simulations, we use $\ell = 1$ and $\ell = 2$). The functions $Q (x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp (-t^2/2) \ dt$ and $Q_1 (\alpha, x) = \int_{-\infty}^{\alpha} t \cdot \exp (-t^2/2) / \Gamma (3/2)$ are the complementary cumulative distribution function (ccdf) of a standard random variable and the Marcum’s Q-function [49], respectively, with $\Gamma (\cdot)$ denoting the modified Bessel function of the first kind and order 0. Note that (33a) is fulfilled by both classes.

Throughout our simulations, the noise power is assumed to be $\sigma^2 = 50$ nW, whereas the maximum power constraint is $p_k = P = 0.5$ W. The distance $d_k$ between the $k$-th transmitter and the access point is uniformly distributed between $3$ and

<table>
<thead>
<tr>
<th>type</th>
<th>coherent</th>
<th>non-coherent (mod-\ell)</th>
</tr>
</thead>
<tbody>
<tr>
<td>identifier</td>
<td>$\rho_k = 0$</td>
<td>$\rho_k = \ell$</td>
</tr>
<tr>
<td>strategy</td>
<td>$u_k = z_k e^{\theta_k}$</td>
<td>$u_k =</td>
</tr>
<tr>
<td>$P_{FA}$($\gamma_k, \lambda_k; p_k$)</td>
<td>$Q \left( \lambda k \sqrt{2 \gamma_k} \right)$</td>
<td>$e^{- \gamma_k^2 / (2 \lambda_k)}$</td>
</tr>
<tr>
<td>$P_D$($\gamma_k, \lambda_k; p_k$)</td>
<td>$1 - Q \left( (1 - \lambda_k) \sqrt{2 \gamma_k} \right)$</td>
<td>$Q_1 \left( \sqrt{2 \gamma_k}, \lambda_k \sqrt{2 \gamma_k} \right)$</td>
</tr>
<tr>
<td>$\Delta$($\gamma_k, z_k; p_k$)</td>
<td>$\frac{2}{\sqrt{2 \gamma_k}} \left( \frac{Q_1(\frac{1}{2} Q(\gamma_k, \lambda_k)^2)}{Q(\gamma_k, \lambda_k)} - \log Q(\gamma_k, \lambda_k) \right)$</td>
<td>$\left( \sqrt{2 \gamma_k}, \lambda_k \sqrt{2 \gamma_k} \right)$</td>
</tr>
</tbody>
</table>

| Table I Mathematical description of practical synchronizers. |
100 m. The channel gains $h_k$ are assumed to be Rayleigh-distributed, with $\mathbb{E}\{h_k^2\} = 1.5 \cdot (d_0/d_k)^2$, where $d_0 = 10$ m is the reference distance between the transmitter and the receiver. The Rayleigh distribution is adopted to emulate the effects of shadowing to some extent.

Figs. 8 and 9 show the behavior of the joint threshold and power control as a function of the iteration step $n$. The results have been obtained using a random realization of the network with spreading factor $M = 64$ and $K = 5$ users. Note that these parameters have been intentionally kept low for the sake of graphical presentation. The users make use of the synchronization strategies $\rho_k = \{0, 0.2, 1, 2\}$, and place the QoS requirements $\mathbf{P}_{FA,k} = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-7}, 10^{-8}\}$. Given this configuration, $\gamma_k = \{6.8, 8.4, 10.6, 12.1, 12.7\}$ dB, and $\gamma_k^* = \{7.8, 9.7, 11.8, 13.3, 13.9\}$ dB. Note that condition (33b) holds, since $\Phi \equiv 0.93 < 1$. In this snapshot, $h_k = \{0.19, 11.55, 0.10, 0.18, 0.12\}$. According to (37) and (38), $\lambda_k^* = \{0.89, 0.86, 0.76, 0.87, 0.75\}$ and $p_k^* = \{1.6, 6.8e4, 13.3, 5.4, 14.2\}$ $\mu$W. As can be seen, the terminals rapidly converge to the desired pairs $(\rho_k^*, \lambda_k^*)$, with a convergence rate of the algorithm proportional to $1/(1 - \Phi)$. At the beginning of the procedure, $\lambda_k^* > 1$ for some users. This means that the QoS constraint is not met. As a consequence, such terminals cannot apply the desired synchronization strategy (their detectors are switched off), and they remain in an initial phase of power control only. However, after a few steps, $\lambda_k^* \leq 1$, which yields $P_{FA,k} \leq P_{FA,k}$. This can be easily verified in Fig. 10, which reports the behavior of $P_{FA}$ as a function of the iteration step. In this network configuration, after the fourth step all users meet their QoS requirements, while maximizing $P_D$ (Fig. 11).

VI. SUMMARY AND CONCLUSION

After a broad introduction to the topic of game-theoretic resource allocation in wireless communication, this paper investigated the issue of game-theory-based criteria to optimize the function of initial code acquisition in a CDMA network. Using the tools of game theory, the problem was restated as a noncooperative (distributed) game in which the transmitter-receiver pairs (the players) jointly set their transmit
powers (at the transmitter side) and detection thresholds (at the receiver side) to maximize the ratio between the probability of detection and the transmitted energy per acquisition. Users are also supposed to place QoS requirements in terms of maximum probability of false alarm and to choose their preferred synchronization strategy (coherent, noncoherent) based on the possible knowledge of some a-priori parameters. General properties of the Nash equilibrium of the game were investigated, also deriving explicit expressions for the strategies at the equilibrium as functions of the network configuration, and sketching a possible implementation for a distributed algorithm.

The main conclusion is that the Nash solution for energy-efficient distributed synchronization based on maximum detection probability per consumed Joule of energy can be found following a criterion that is similar to the one used for energy-efficient data detection based on maximum throughput per Joule. Nonetheless, the two cases show significant differences when applied to the same CDMA scenario, leading to different values of optimum transmitted power by the mobile terminals. The two solutions may be not in conflict in a network set-up scenario, since the criterion for optimum synchronization will be used in a first instance, to revert to the data-detection optimal criterion after synchronization is over. Further research is needed to assess the case of a mixed population, i.e., some terminals already in-sync and using the data-optimization criterion co-existing with other terminals still in the acquisition phase.

REFERENCES

Giacomo Bacci (S’07, M’09) received the B.E. and the M.E. degrees in telecommunications engineering and the Ph.D. degree in information engineering from the University of Pisa, Pisa, Italy, in 2002, 2004, and 2008, respectively. Since 2005, he has been with the Department of Information Engineering at the University of Pisa, where he is currently a post-Doc fellow. In 2006-07, he was a visiting student research collaborator at the Department of Electrical Engineering at Princeton University, Princeton, NJ, USA. Since 2008, he has also joined Wiser srl, Livorno, Italy, as a software engineering. Dr. Bacci’s research interests are in the areas of digital communications, signal processing and estimation theory. His current research topics focus on resource allocation for multiple-access wireless networks, time delay estimation for satellite positioning systems and wireless communications, and channel coding for IMT-advanced technologies.

Marco Luise (M’85, SM’94) is a Full Professor of Telecommunications at the University of Pisa, Italy. After receiving his M.E. and Ph.D. degrees in electronic engineering from the University of Pisa, he was a Research Fellow of the European Space Agency (ESA) at ESTEC Noordwijk, The Netherlands, and a Researcher of the Italian National Research Council (CNR), at the CSM/DR Pisa. Prof. Luise co-chaired four editions of the Tyrrhenian International Workshop on Digital Communications, was the General Chairman of the URSI Symposium ISSEP’98, and the General Chairman of EUSIPCO 2006 in Florence, Italy. He will be the General Co-Chairman of EW 2010, to be held in Lucca, Italy, and of IEEE ICASSP 2014, to be held in Florence, Italy.

Prof. Luise served as an Editor for Synchronization of the IEEE TRANSACTIONS ON COMMUNICATIONS, and as an Editor for Communications Theory of the European Transactions on Telecommunications. He is the co-Editor-in-Chief of the International Journal of Navigation and Observation, and acts as the General Secretary of the Italian Association GTTI, Gruppo Telecomunicazioni Teoria dell’Informazione. He is also the Vice-Chairman of the URSI’s Commission C and a member of the International Committee on Global Navigation Satellite Systems (ICG) of the UNO. Prof. Luise has authored more than 200 publications on international journals and contributions to major international conferences, and holds a few international patents. His main research interests lie in the area of wireless communications, with particular emphasis on CDMA/multicarrier signals, software-defined radio, satellite communications and positioning.
Abstract—Network opportunism in wireless systems aims at jointly exploiting the resources of separate networks. So far, small emphasis to this paradigm has been given in the literature related to Wireless Sensor Networks (WSNs). This paper first describes a vision on the evolution of WSNs towards the application of this concept. Then, a way to allow exchange of information related to the available resources among networks, is formalised. Finally, the concept is exemplified by considering a specific type of application scenario and resource sharing approach: a vehicular network where pairs of mobile nodes exchange packets by exploiting a store-carry-forward mechanism. This scenario allows the introduction of the concept of Sociability Based Routing.

I. INTRODUCTION

NETWORK opportunism in wireless systems is a recent concept, aiming at jointly exploiting the resources of separate networks according to the needs of specific application tasks [1]. So far, emergency scenarios, and Delay Tolerant Networks (DTNs) based on MANETs (Mobile Ad Hoc Networks), have been mainly discussed in the context of Opportunistic Networks (ONs) [2], [3], with small emphasis on the role of Wireless Sensor Networks (WSNs).

This paper first describes a vision on the evolution of WSNs towards the Internet of the Environment, discussing the concept of network opportunism in WSNs (Environmental Opportunistic Networks, EONs). Then, the paper provides a first attempt to formalise a way to allow exchange of information related to the available resources among network nodes, and networks (either WSNs or of other type), a feature necessary to the development of the paradigm of ONs. Finally, the concepts are exemplified by considering a specific type of application scenario and resource sharing approach, which in the past few years attracted the attention of researchers in the area of DTNs: a vehicular network where pairs of mobile nodes exchange packets by exploiting a store-carry-forward mechanism. This scenario allows the introduction of the concept of Sociability Based Routing, which is discussed in this paper. This routing strategy selects a subset of optimal forwarders among all the nodes and relies on them for an efficient delivery [5], [6].

II. EVOLUTION OF WSNs TOWARDS THE INTERNET OF THE ENVIRONMENT

Wireless Sensor Networks (WSNs) are composed of nodes able to cooperatively sense the environment, and possibly control it through actuation devices [4]. Different tasks can be accomplished by a WSN, depending on application needs and the type of physical entities measured/controlled. Future WSNs will be designed so that different application types will be able to run over the same network. Sensing and actuation are respectively the initial and final steps of a process that also involves functionalities like data processing, communication, and possibly storage, data aggregation, etc.

It is expected that in the next few years billions of wireless sensor nodes will be deployed over the globe in various types of environments, with very diverse tasks, hardware, software and communication capabilities. As WSN technologies are becoming more and more evolved, more complex tasks can be accomplished that require convergence of all such functionalities, with their multi-faceted options: nodes can have multiple and diverse sensors/actuators for diverse physical entities to be measured/controlled; communication among nodes can take place by means of separate air interfaces, possibly through infrastructure networks; data processing can require different computing/storage capabilities depending on the type of data to be processed; and so on. On the other hand, to keep the cost of nodes at reasonable levels, they can not be equipped with all such functionalities, at the deepest degree. Therefore, WSNs in the future will be composed of heterogeneous nodes enabling only part of such functionalities, and to diversified extents; each node will be characterised by different resources (sensors, actuation devices, communication, storage and computation capabilities, etc). We denote this new class of heterogeneous WSNs as Differentiated Resource WSNs, to emphasise that the resource types owned by nodes will be the basic feature to differentiate them. As the deployment of WSNs will become increasingly pervasive, evolution towards such class will be a natural fact.

This evolution characterises all types of wireless networks, either environment-related or not, but it will be much more evident for WSNs, owing to the need to keep complexity and cost of devices at minimum levels.

According to the specific task required by the application, the resources available over the network have to be aggregated in order to accomplish the task. So, the network (as a whole) must be equipped with functionalities able to map the availability of resources, and aggregate them in a coherent way, according to application needs. Node mobility, idle/active cyclic states, expiration of nodes, can impact the process of resource mapping, making it more difficult as the network time evolution requires frequent updates.

Suitable methods must be devised to simplify the task of resource mapping, classifying the different types of resources that can be made available and allowing easy exchange of data within network nodes representing the availability of resources.
Differentiated Resource WSNs will operate in environments where several other (non environment-related) wireless networks will be available, such as GSM/UMTS, WiFi hot spots, etc. If gateways will be deployed, letting networks using different communication standards inter-operate, the separate WSNs will form an heterogeneous Internet of the Environment (different from the Internet of Things, where objects cooperate with no special emphasis to their possible roles as sensing devices) spreading all over the globe (and outside). Gateways will be specifically deployed for this purpose in some cases. However, the use of gateways not specifically deployed to such aim, roaming the environment randomly, will be much more convenient: mobile radio users, while travelling with their mobile phone terminals, laptops, palmtops, connected to infrastructure networks like GSM, UMTS, LTE, etc, will act as gateways connecting the Internet of the Environment to the Internet (of Computers) and the so-called Internet of Things based on ubiquitous computing and radio frequency identification of objects, a concept developed at ITU level.

Every WSN will then be one single component of an heterogeneous aggregation composed of networked objects and sensed environments, with mobile roaming gateways connecting them to the Internet through infrastructure networks.

In this context, the mobile gateways play a fundamental role, permitting connectivity among separate environmental islands. As they are carried by people, the human space becomes central in this architecture. However, it is worthwhile stressing that humans in this context are not (necessarily) the final users, and just become means to let the environment be connected to all other networks. So, this is more a people-centric, rather than human-centric vision: the individuals do not play a role, they are meaningful only as carriers of information.

III. OPPORTUNISM IN WSNs

In some application scenarios, it might happen that multiple parallel tasks are accomplished by the WSN (for instance, in the case that publish/subscribe paradigms are implemented, and multiple subscribers operate simultaneously). The multiple parallel tasks compete for network resources, and the latter are exploited opportunistically by the single tasks according to their needs, and their (possibly limited) knowledge about resource availability. This is a form of intra-network opportunism.

It might even happen that the resources available in the seed WSN are not sufficient to accomplish the task; in such situation, the network can try to scan the radio environment, detect the presence of other networks deployed for different tasks (e.g. WiFi hot spots, or computer networks in an office environment, or GSM/UMTS public networks) and address such helper networks trying to exploit their available resources. This requires ability to monitor the radio environment to find candidate helpers, and to communicate with them through different air interfaces. This type of ONs is a new paradigm proposed recently by Lilien et al. [1], even if it was not applied by the Author to the specific case of WSNs.

A peculiar type of network opportunism has been recently investigated by some authors, such as Conti et al. [2], [3], where information is routed from source to destination in mobile ad hoc networks (MANETs) through nodes that allow exchange of data even in the absence of a complete path, taking advantage of their mobility that allows interconnection between separate islands of nodes, at the cost of an increased delivery delay. This form of opportunism can be considered as a specific case with respect to Lilien’s definition, if mobility itself is considered a resource provided by a separate helper network. This type of ON also represents a sub-class of the DTNs.

If node mobility is considered as one of the functionalities possibly provided by nodes, and as such it is a network resource, then both types of network opportunism can be treated from a formal viewpoint in a coherent way, where a seed WSN exploits resources of helper networks. Section V of this paper exemplifies such situation.

However, the case where the WSN is the seed of this opportunistic process, looking for helper networks according to specific needs, does not represent the most interesting one. In fact, this requires implementation of agents/entities in the WSN able to detect candidate helpers and run decision processes, a requirement that might be difficult to fulfill in some cases owing to the need to keep costs and complexity at reasonable level. A situation where other types of networks, implementing more complex devices, might need the sporadic help of WSNs deployed in the environment, can be more frequent and relevant. In these cases, the seed network (for instance, a MANET or a WLAN) might receive the command to accomplish an urgent task, generated by an unexpected event, and might scan the environment to check whether some help can be found from WSNs deployed in the environment. A sample scenario is reported later.

In all cases, either if the WSN is the seed or the helper of the opportunistic process, the aggregation of different types of networks can be seen under a unified view, classifying node resources according to a common formalism; inclusion of the role played by WSNs, requires consideration of the presence of functionalities as sensing and actuation, and the need to handle very heterogeneous deployments. This paves the way to the new paradigm of Wireless Environmental Opportunistic Networks (EONs), where sensing, communication, data processing and storage, actuation, control, mobility, gateway capabilities are all functionalities (resources) made available by nodes distributed in the environment, with the objective of sensing it, and possibly control it through direct actuation or the interaction with human users.

IV. THE RESOURCE VECTOR FOR EONs

Let us now try to provide a formalism able to classify the different types of resources that can characterise nodes in a EON.

The aim of this classification is to define a simple method for comparing candidate helpers; it is not a precise definition of the potential help provided, as this would require lots of specific information.

We distinguish the following types of node functionalities that may be useful for the accomplishment of an application
task in EONs, and as such can be considered as network resources: Sensing, Communication, Processing, Repository, Decision, Sociability, Gateway, Actuation.

- Sensing (denoted by its initial S in the following) is a multi-faceted functionality, as there might be different types of sensors deployed. Assuming a maximum number N of sensors can be distinguished (for temperature, acceleration, pressure, etc), S for a specific node is a vector of N binary values, where 1 represents presence, and 0 absence of the specific sensing capability.

- Communication (C in the following) is a functionality provided by all nodes of a wireless network. However, nodes can have different communication capabilities. C is measured in Mbitm2/s in the following, putting together the (application level) data rate and the coverage area. To normalise C to one, we consider as reference maximum value 109. Any value for C is represented as compared to the reference value. For instance, C = 0.1 for a node represents a value of 108 Mbitm2/s.

- Processing (P in the following) is a functionality provided by all nodes to very different extents. P is measured in Mips. To normalise P to one, we consider as reference maximum value 103.

- Repository (R in the following) is a value representing the capability of a node to store data. R is measured in MBytes. To normalise R to one, we consider as reference maximum value 103.

- Decision (D in the following) is a value representing the capability of a node to take decisions after the data has been taken from the environment through the sensors. The decision can be taken automatically, providing inputs to actuators, or by interacting with human users through specific interfaces. D is a binary value where 1 represents presence, and 0 absence of the capability to take decisions.

- Sociability (B in the following) is a time-varying scalar parameter that has to do with the frequency and type of nodes encounters over a certain time window (see Section V).

- Gateway (G in the following) represents the ability of a node to act as gateway towards other networks using separate air interfaces. Assuming the presence of K separate networks, G is a vector of K binary values, where 1 represents presence, and 0 absence of the specific capability to act as gateway. One value is set to 1 as default, the one representing the air interface used within the network;

- Actuation (A in the following) is a multi-faceted functionality, as there might be different types of actuators deployed. Assuming a maximum number I of actuator types, A for a specific node is a vector of I binary values, where 1 represents presence, and 0 absence of the specific actuation capability.

Given this formalism, every node in an EON can be represented by a vector of size N + K + I + 5. Such vector, denoted as Resource Vector in the following, is mostly static and represents the set of resources made available by the node. The Resource Vector needs to be exchanged with other nodes.

The potential resources of a sub-network of an EON can be measured through a Network Resource Vector given by the sum of the Resource Vectors of all participating nodes. This Network Resource Vector is a concise element that can be published by (some) nodes of a wireless network, either WSN or of any other type (cellular, etc), to easy the implementation of the paradigm of EONs. Through the detection of nodes publishing the Network Resource Vector, a seed network can decide whether a candidate helper is suitable, according to the needs of the specific task to be accomplished. Suitable comparison between the Network Resource Vector of the candidate helpers and of the seed network, can provide indication of which resources are available overall, and what should be selected opportunistically.

Now, the problem is how to aggregate in a simple way the information coming from several nodes in a sub-network, to generate the elements of the Network Resource Vector. Clearly, some resource types (like processing) do not sum up, since the overhead to synchronise the sub-tasks among the various nodes is such that the overall resource (obtained gathering the resources from the different nodes) is smaller than the sum. In other cases, the opposite is true. In particular, Sociability is a property which is depends in a very complex way by the interaction between moving nodes. Therefore, this is one of the elements of the Resource Vector which are more complex to define and handle in an aggregate way. This will be further elaborated and exemplified in Section V.

V. Sociability Based Routing in a EON Scenario

In the present section we aim at considering a specific scenario where mobile nodes take forwarding decisions based on the Sociability parameter (B) of the resource vector, resulting in Sociable Routing, a novel routing scheme for ONs. Although illustrative, this is not meant to be exhaustive of the potential benefit of using the Resource Vector in EONs.

The key idea of Sociable Routing is to solve the routing problem in DTNs [5] by assigning to each network node a time-varying scalar parameter depending on its social behavior, called sociability indicator, that has to do with the frequency and type of node’s encounters. Then, each node forwards its data packets only to the most sociable nodes. Thus, the chances of reaching the intended endpoint are maximized and the amount of transmissions kept under control. There are some analogies between this approach and the game theoretic concept of player’s reputation adopted to model security in wireless networks [7].

After giving a detailed formalization of the sociability concept, we simulate packet transmissions in a DTN in an urban context. In particular, we consider a case study where nodes are vehicles moving according to real traffic traces [8].

A. Sociability Concept

The basic idea is that nodes having a high degree of sociability (i.e., frequently encounter many different nodes) are good candidate forwarders. Applying this simple rule to a delay tolerant network is quite straightforward. As first step, one needs to observe nodes behavior and learn their habits.
Then, a synthetic scalar parameter shall be assigned to each node depending on its social behavior. Finally, routing from a source to a destination node is performed by forwarding packets to a restricted set of relays which show a high degree of sociability and, thus, are very likely to get in touch with all possible endpoints.

One further assumption that we need is the periodicity of behaviors, meaning that it is possible to make predictions on the social conduct of a node based on what has been observed before. Roughly speaking, we expect those nodes that showed very high sociability over a time period of a certain length to behave accordingly in the future for a period of at least the same length. This is a reasonable hypothesis in population networks, and we believe it still is in all scenarios where the mobility of nodes is governed by human behavior, as in vehicular networks, pedestrian networks, etc.

1) Modeling: The way in which the social features are modeled should be very simple, on the one hand, in order for the nodes to produce and exchange such information in an inexpensive manner. On the other hand, the challenge stands in capturing as much as possible of the exploitable information in a single parameter, that we shall call sociability indicator.

One way sociability could be quantified is by looking at the intercontact information of each node [9]. In particular, the intercontact time analysis reveals how frequently a node meets with one another. As an example, an indication on the average intercontact time of a node with any other could give a rough idea of its social behavior. However, in the latter case, one can appear very sociable by having frequent meetings with a very restricted set of neighbors. Unfortunately, this does not make it a good candidate forwarder.

Moreover, an important aspect to be captured in analogy with human relationships, is that one person who only meets a single friend, the latter being very sociable, can itself be considered sociable. Turning to an information network perspective, a node being isolated most of the time with very sporadic links to a single neighbor, may appear very unsociable. Nonetheless, if the neighbor is very sociable and can reach many destinations, then the former node may also have chances to send its packets to many destinations through a 2-hop path. As a consequence, the presence of sociable neighbors is an important addendum that should be incorporated into the sociability indicator of one node.

Intuitively, it is a natural assumption that mobility patterns of nodes are related to their social behavior. In fact, if a node visits a great number of different locations in a short time, it is likely to meet many others. Although this is true to some extent, there are plenty of scenarios where the concentration of users is not constant in space (e.g., the union of a city center with its suburbs). Hence, the mere covering large distances does not necessarily result in high forwarding opportunities. For this reason, in order to maintain the overall idea detached to any specific environment, we chose not to include any direct information regarding mobility patterns in the sociability indicator.

In [11], the Authors state that two people having similar mobility patterns (in terms of frequency of visits to specific locations) are more likely to meet each other, thus to be able to communicate. Then, they recognize that the main limitation of the previous statement is that even though two people visit the same locations, they do not necessarily do it synchronously. Thus, two such nodes may never be in the range of each other. This is not a rare event, especially at urban scale. Consider for example a public transportation fleet (e.g., buses). Two buses running on the same route have the exact same mobility patterns. However, if one follows the other few kilometers behind, they never reach each other. More generally, there are places like, e.g., a big mall, that many people periodically visit at different time. This results in some similarity of their patterns which does not necessarily reflect meeting opportunities.

Finally, we emphasize that the sociability indicator only highlights what are the best forwarders in a given time period, in the sense of those having the highest degree of sociability. As a consequence, this information is not related to a specific destination to be reached but it is instead absolute. This descends from avoiding a sociability characterization based on mobility patterns and is consistent with the intent of minimizing the exchange of data. This also implies that no prior knowledge of the destination (e.g., its position, sociability indicator, etc...) is requested at the source.

2) Acquisition: Since we do not use information on positions, nodes are not requested to adopt any positioning technique, nor do they have to learn their mobility patterns as in [11]. The two main issues arising with the use of Sociable Routing are i) how a node learns its own social behavior and ii) how it communicates its social behavior to other nodes.

Note that the two issue are strictly connected, as a node needs to know the social behavior of its neighbors in order to derive its own. For this reason, a distributed strategy where nodes, upon encounters, update their own sociability parameter through the exchange of a minimum amount of data, could be the optimum. For example, the sociability updates could be appended to data bundles in order not to overwhelm the network with signaling information. However, this is not addressed here, since our aim is primarily that of presenting and validating the general idea at the base of Sociable Routing.

Hence, in the following we assume that nodes have knowledge of their social behavior referred to a specific time window. In a practical setting, this is equivalent to assuming that: i) a node keeps track of the contacts with its immediate neighbors; ii) it sends this information to a central processing unit (which could be a node itself) that combines it with that of all the other nodes; iii) the central processing unit broadcasts a vector containing the updated sociability indicators.

3) Usage: As previously mentioned, the basic idea is to select a set of sociable nodes that can potentially reach any endpoint. This set should be kept small enough to avoid useless transmissions. To this end, the following strategy can be adopted. A node takes its routing decision at a given time \( t \) by i) evaluating the sociability indicators of the current neighbors; ii) comparing them to its own sociability indicator and iii) choosing as forwarders a maximum of \( N_f \) nodes that have greater sociability than its. This simple scheme allows to limit the number of bundle transmissions at each encounter by setting a maximum, \( N_f \). Moreover, a node does not transmit
any bundle if it does not meet any more sociable node. As a further implication, when a bundle is generated by a node with low sociability degree, a large number of transmissions are permitted, since the source will certainly meet more sociable nodes. On the contrary, if the bundle is generated by the most sociable node, there will not be any transmission until the source is itself in the range of the destination, since it is also the best possible forwarder. This seeming imbalance is explainable as follows. Because an unsociable source is likely to remain isolated for a long time, it makes sense for the network to put a greater effort to route its message along by generating replicas. In the opposite case, when a source is highly sociable, only few transmissions are required because mobility will do the rest.

In a formal tone, by using a notation similar to that of [11], let $U$ be the set of all nodes and $N = |U|$ their number. The sociability indicator of a node $k \in U$ at time $t$ is $s_k(t) \in [0, 1]$. Assume also that at time $t$ node $k$ has a number of active direct links to some neighbors. Let us denote as $W_k(t) \subseteq U$ the neighborhood of $k$. The routing decision of $k$ consists of either keeping the bundle or selecting up to $N_f$ next forwarders belonging to $W_k(t)$. With respect to a destination node, $b$, this can be performed by using a decision algorithm to be applied to the set $W_k(t)$ and $b$, and yields the set, $R_k(t) \subseteq W_k(t) \subseteq U$, $|R_k(t)| \leq N_f$, of next forwarders. The pseudocode is given in Algorithm 1.

**Algorithm 1 Routing decision algorithm**

1: $R_k(t) := \emptyset$
2: if $b \in W_k(t)$ then
3:   $R_k(t) = \{b\}$
4: else
5:   $i = 1$
6:   while $W_k(t) \neq \emptyset \cap i \leq N_f$ do
7:     $h := \arg \max_{j \in W_k(t)} s_j$
8:     $W_k(t) \leftarrow W_k(t) \setminus \{h\}$
9:     if $s_j > s_h$ then
10:        $R_k(t) \leftarrow R_k(t) \cap \{h\}$
11:   end if
12:   $i \leftarrow i + 1$
13: end while
14: end if

**B. Simulation results**

In the present section we introduce the simulator that allows us to test the forwarding scheme proposed and to compare it to other existing protocols.

1) Methodology: We have designed an autonomous network simulator for testing the routing scheme. It takes as input a mobility trace and generates mobile nodes accordingly. The time is discretized and resolution is 1 sec. Each node has an infinite buffer for storing the exchanged packets. In a realistic setup, a routing protocol should be evaluated by accounting for limited buffering capabilities. Nonetheless, although we do not address it here, we assess the validity of protocols by also counting the amount of extra packets generated, as a rough measure of resources consumption at network level.

In addition, we make very simple assumptions at physical and MAC layers, namely, nodes are in contact when their distance is less than the transmission range, $TR$; channels are interference-free; and transmissions are instantaneous. Furthermore, although a node is not aware of its absolute geographical position, it has a complete knowledge of its logical connectivity, (i.e., what other nodes are within its transmission range), and it is always willing to cooperate with others.

A simulation run starts when two nodes are randomly selected as source and destination of a bundle, respectively, and terminates when the bundle is either successfully received by the recipient or discarded for exceeding a timeout threshold.

2) Input mobility and parameters: As input mobility, we consider the taxi cab traces available in [8]. It must be noted that taxi cabs’ movements are not particularly predictable as can be those of a private citizen or even a public transportation vehicle (e.g., a bus). In fact, apart from the most frequent routes (e.g., airport to train station), each time a passenger is collected, a destination which potentially differs from the previous one has to be reached. For this reason, if we can appreciate any benefit from the Sociable Routing scheme in this scenario, we expect even better performance when using, e.g., Seattle city bus traces [10] as input mobility. However, this comparison is left for future work.

We put two constraints in order to speed up the simulations. First, source and destination nodes are randomly picked among those that are located, at the generation instant, in a 10 x 10 km square centered in downtown San Francisco. This indeed decreases the average delivering time by avoiding too far away source-destination pairs. Secondly, nodes that have not been moving for more than 1 hour cannot be source candidates. This avoids extra delays due to when a packet is generated by a cab that is not in service, and thus has greater chances to remain isolated for long.

The number of nodes, all included, is then 535 and the traces are two weeks long. Every simulation is composed of 1000 runs (i.e., 1000 bundles are either successfully received or dropped due to excess delay) and is started at a random time on the first day of traced period. We set a timeout of 1 day and a transmission range $TR = 500$ meters. This value is in accordance, for example, with the standard IEEE 802.11p [12], which is meant to be employed in vehicular networks. Finally, in case of multiple contemporaneous encounters, one node is allowed to forward the bundle to only $N_f = 1$ neighbor.

3) Results: The performance of our routing scheme, Sociable Routing, is compared against that of other known protocols, namely, Epidemic [13], MobySpace [11] and Random. In the latter, packets are forwarded independently of nodes degrees of sociability.

When simulating Sociable Routing, the time interval between two refreshes of the sociability indicators must be set. This should be calibrated based on the nature of mobility traces. We assume no a-priori information is available about the social behavior of the nodes. We then take $T = 1000$ sec as initial guess.
**REFERENCES**


Trends in Adaptive Modulation and Coding
Andreas Zalonis, Natalia Miliou, Ioannis Dagres, Andreas Polydoros, and Hanna Bogucka

Abstract—This paper presents a review of techniques proposed in the literature that target the best use of resources available in a wireless communication system. These are referred to as Adaptive Modulation and Coding (AMC) techniques. A brief overview of fundamental communication limits, in particular channel capacity, is also provided in order to establish the limits for any adaptation algorithm. Furthermore, an appropriately chosen example is presented in order to demonstrate the usefulness of accurate performance modeling in the AMC design. Finally, challenges and future problems are mentioned.

Index Terms—Index TermsAdaptive, modulation, coding, capacity, OFDM

I. INTRODUCTION

In recent years there has been a significant increase in data-rate requirements described in the standards of new and upcoming wireless communication systems. In order to increase the data rates offered, a simple approach is to increase the allocated bandwidth. This approach is expensive, since the radio-frequency spectrum is nowadays fragmented and occupied by a plethora of systems. Therefore, research over the last years has been focused towards improving spectral efficiency, so that higher data rates can be achieved within a given bandwidth. Targeting to the inherent capacity of the underlying channel, techniques which adapt and adjust (in real-time) transmission parameters based on the link quality have been proposed. These are collectively referred to as “Adaptive Modulation and Coding” (AMC) and they provide as their output the values of transmission parameters to be employed in a following transmission period, based on feedback information and in accordance with particular cost functions related to the targeted Quality of Service (QoS).

In this paper, a review on the wireless-channel capacity definitions and related issues is first presented, with an emphasis on communication scenarios that involve the existence of Channel Side Information (CSI). This is of great importance for AMC since it provides the fundamental limits for its performance. In the subsequent section, the AMC framework is presented with an emphasis on practical considerations for algorithmic design in the context of multi-modal operation, along with a literature review of the proposed techniques. In Section IV, an example based on the discussed framework is presented which demonstrates the need for accurate compact link level modeling in AMC design for OFDMA systems. In Section V some challenges for future research conclude the paper.

This work was supported by the EC in the framework of the FP7 Network of Excellence in Wireless Communications NEWCOM++ (contract no. 216715). Andreas Zalonis, Natalia Miliou, Ioannis Dagres and Andreas Polydoros is with the National Kapodistrian University of Athens, Greece (e-mail: {azalonis, nmiliou, jdagres, polydoros}@phys.uoa.gr). Hanna Bogucka is with the Poznań University of Technology, Poland (e-mail: h bogucka@et.put.poznan.pl)

II. FUNDAMENTAL COMMUNICATION LIMITS

In this section a brief survey of schemes that model the existence of multiplicative noise (fading) in a wireless environment is provided, since such fading is the main source of channel variability that dictates the need for AMC methods and techniques. Furthermore, the fundamental information theoretic limits of such schemes are characterized. The main goal here is neither to provide an exhaustive analysis of the topic, nor to explain in depth the existing techniques. Instead, we aim to highlight the potential benefits of AMC.

A. Channel Capacity

The evaluation of capacity bounds for various channel models and system scenarios has been an interesting research topic since Shannon’s pioneering work. The determination of such performance limits defines the framework for the design and development of optimal communication techniques. In many cases, the computation of the exact value of channel capacity is a difficult problem. In some cases, these bounds can lead to the exact determination of channel capacity. In all cases however, any available analytical characterization of a system’s information capacity serves as a performance criterion and a tool for the system design.

B. Channel Capacity of Fading Channels with Channel Side Information

In wireless transmission, channel conditions may vary arbitrarily due to changes in the fading environment or due to the users’ mobility. In most practical cases, and depending on the relative dynamics, the system is able to extract valuable information regarding the prevailing level (strength) of the channel-fading process. This type of information is usually referred to as CSI and can either be available at both ends (the transmitter as well as the receiver), or at the receiver (Rx) only but not at the transmitter (Tx). The degree of accuracy of this CSI is certainly a critical parameter for the performance of the system under consideration. The case in which the channel is assumed to be perfectly known at the receiver and/or the transmitter has been studied extensively in the literature. In a Rayleigh-fading Single-Input Single-Output (SISO) setting, it was addressed in an early work by Ericsson [1], where analytical expressions for the capacity of flat-fading channels with perfect receiver CSI have been derived. In a more recent work (Ozarow et al. [2]), results were derived for the average as well as the outage capacity in cellular mobile radio, assuming perfect CSI at the Rx.

In recent OFDM-type designs, a system implements a strategy whereby the estimated values of the CSI are relayed to the Tx with different levels of accuracy. It is therefore of
particular interest to re-examine past results on the capacity of fading channels but now with Tx-side information available. Goldsmith and Varaiya [3] analyzed the capacity of flat fading channels with perfect CSI at the transmitter and/or the receiver. Borade and Zheng [4] investigated, among other scenarios, the channel capacity in the low signal-to-noise-ratio (SNR) regime when both sides have perfect CSI. They showed that, for very low values of the SNR the capacity is $\text{SNR} \log(1/\text{SNR})$ and this is achieved by on-off signalling with a fixed “on” level.

In practice, however, true access to perfect CSI cannot be guaranteed to either side, mainly due to the rapid changes in the fading environment and limited energy that can be allocated to non-data carrying symbols to assist the channel estimation. Thus, extensive research has addressed the case where CSI is either imperfect or, in the other extreme, completely unavailable to either side. Abou-Faycal et al. [5] have studied the capacity of the Rayleigh fading channel with purely unknown fading levels. They showed that the optimal input distribution is discrete. This is a somewhat surprising result, especially when compared with the optimality of the continuous Gaussian input distribution when CSI is available and perfect. Under a peak constraint on the input of a Rician fading channel, this discreteness property of the capacity-achieving input distribution has also been proven in [6]–[8] and, more generally, for a broad class of SISO channels in [9].

In a MIMO setting the capacity and optimal input distribution of a Rayleigh fading channel have been derived by Marzetta and Hochwald [10] where CSI is also unavailable to both sides. This work triggered further research on the discreteness of optimal input distributions on MIMO settings. Zheng and Tse [11] addressed the capacity of MIMO Rayleigh fading channels in high SNR.

In most cases of interest, however, these extremes, of having either perfect CSI or no CSI at all, are not valid. In particular, practical OFDM systems tend to lie between those two extremes, except when there is very high mobility. The analysis of the capacity of fading channels with imperfect or “noisy” CSI is therefore of great practical interest. Médard [12] investigated the effect of imperfect channel knowledge on the channel capacity and obtained upper and lower bounds on the achievable mutual information rates. Lapidoth and Shamai [13] analyzed the effects of channel estimation errors on performance whenever Gaussian codebooks are employed along with nearest-neighbor decoding. The capacity of imperfectly known fading channels is addressed in [14] for the low-SNR regime and in [15] for the high-SNR regime. These results, however, have not considered explicit training and estimation techniques and the needed resources allocated to do so. An analysis of channel capacity in a training-based communication setting with Rayleigh block-fading can be found in [16]. Hassibi and Hochwald [17] looked into some training schemes for multiple-antenna channels recently.

We shall now focus further on fading channel capacity with CSI available at the Tx. When CSI concerns channel fading values, CSI is treated as causal side information at the Tx (as opposed to noncausal side information that is not treated in this survey). In this case, techniques such as adaptive rate/power control, MIMO beam-forming, water-filling etc. are all applicable. The causal case of Tx-CSI was first introduced by Shannon [18] wherein he showed that the CSI-endowed channel can be transformed into a no-CSI channel of an exponentially larger alphabet size.

Assuming causal Tx-CSI, we now describe briefly the results derived for typical fading models in the literature.

**Slow fading:** When the Tx knows the CSI, one option is to control the transmit power such that the corresponding full information rate can be delivered regardless of the fading state. Regarding the power this effectively “channel inversion” strategy guarantees a constant receiver SNR, irrespective of the channel gain. Regarding the rate adaptation some pre-specified rates are chosen when such exact channel inversion is feasible. However, a very large amount of power must be spent in order to ‘invert’ a very bad channel, which encounters practical limitations of peak-power-constrained transmission.

**Fast Fading:** The goal now is to maximize the average information rate where the averaging occurs over many coherence-time periods. The optimal power and rate allocation in this case is based on the water-filling principle. In general, the Tx allocates more power when the channel is good, taking advantage of this improved channel condition, and less or even nothing when the channel is poor. This is conceptually the reverse of the channel-inversion strategy above. The natural implication of the water-filling capacity is a variable-rate coding scheme.

In the above setting, water-filling is done over time. A duality exists with a frequency-selective channel, where water-filling is done over the OFDM sub-carriers. In both cases the problem can be viewed as that of a bit-power allocation over parallel channels.

Comparing the channel capacity in the case of full CSI at the Tx with the one in the Receiver-Only CSI (RxO-CSI) case, some conclusions can be drawn. In particular, at low SNR, the capacity with full CSI is significantly larger than the RxO-CSI capacity, whereas at high SNR the difference between the two tends to zero. Over a wide range of SNR, the gain of the water-filling procedure over the RxO-CSI capacity is very small. A comprehensive survey of information theoretic results on fading channels can be found in [19]. With imperfect channel knowledge at the transmitter, the capacity is $\beta \text{SNR} \log(1/\text{SNR})$, where $\beta$ is a scalar parameter ($0 < \beta < 1$), describing the fraction of channel energy in the part of the channel known to the transmitter. An analysis of partial transmitter knowledge over Rician channels can be found in [6], [7], too.

**C. MIMO Capacity**

In the above setting, the introduction of multiple antennas under suitable conditions provides an additional spatial dimension for communication and yields gains in degrees of freedom. This results to an increase in capacity: In fact, the capacity of such MIMO channels with N transmit and receive antennas is proportional to N. MIMO communication is a broad and interesting topic with many applications. In particular, in the high-SNR regime, MIMO techniques become the primary tools to increase capacity significantly through the
degree-of-freedom gain previously mentioned, as well as the induced power gain.

The use of multiple transmit and receive antennas provides many benefits for both fast and slow fading channels. In fast fading, antenna diversity introduces a power gain as well as a degree-of-freedom gain. The analysis of fast-fading MIMO is simpler and addresses mainly channel capacity, whereas the analysis of slow fading is generally more complex. In this case, the outage probability is the goal as a function of the target rate. The outage probability reveals in principle the tradeoff that exists between the error probability and the data rate. In slow fading, there is a triple gain from the introduction of multiple antennas, namely in power, degrees of freedom and diversity. In the high-SNR regime, there is an approximation of the outage probability that captures the benefits of MIMO communication for slow fading channels. This is the fundamental tradeoff between the increased data rate (via an increase in the spatial degrees of freedom the multiplexing gain) and the increased reliability (via an increase in the diversity gain). The optimal diversity-multiplexing tradeoff is used as a benchmark in comparing the various space-time schemes and is helpful for the design of optimal space-time codes.

Considering the time invariant Gaussian MIMO channel, the spatial dimension plays the same role as the time and frequency dimensions in the time-varying fading channel with full CSI and the time-invariant frequency-selective channel. The capacity is therefore obtained by a water-filling power allocation scheme, albeit the water-filling takes place in the spatial domain. It depends highly on the singular values of the channel gain matrix, corresponding to the eigen-modes of the channel, or eigen-channels. For high SNR, where the level of “water” is low, it is asymptotically optimal to allocate equal amounts of power on the nonzero eigen-modes. The number of spatial degrees of freedom represents the dimension of the transmitted signal as modified by the MIMO channel. It provides a crude measure of the capacity of the channel. At a low SNR, the optimal policy is to allocate power only to the strongest eigen-mode. In this regime, the rank of the channel matrix is less relevant for the characterization of the channel capacity. Instead, the energy transmitted through the channel is a more critical parameter. For a detailed analysis of MIMO communication as well as the related concepts the reader is referred to [20].

III. COMMUNICATION TECHNIQUES FOR APPROACHING THE LIMITS

A. AMC Framework

As mentioned before, the objective of AMC is to optimize the use of available resources (system bandwidth, channelization, transmit power, time slots, computational power of executing platform) in order to achieve a specific Quality of Service (QoS). The AMC system is composed of: (a) an adaptation criterion or cost function, generally related to the QoS parameters; (b) the hypothesized nature of the CSI that the transmitter needs to know about the channel, which information is often imperfect, erroneous or obsolete; (c) the particular optimization algorithm chosen or developed for this problem; and (d) the resulting outputs of the AMC optimization algorithm (namely, a set of transmission parameters to be employed in the next transmission). As far as the adaptation criteria are concerned, there are multiple possibilities and combinations of AMC schemes in which some particular metrics and quantities may constitute the inputs (requirements and constraints) for one strategy or the criterion/output for the other. Some quantities (such as data throughput) may be introduced as the criterion (objective function), while others (like transmission power limit and delay) may be introduced as constraints. By varying these choices a large number of adaptation strategies are obtained. Accommodating that plethora of options is challenging and, in some cases, may result in very complex solutions that are hard to implement.

An illustrative example of possible choices for cross-layer optimization in a cellular system is depicted in Fig. 1: Tx parameter optimization procedures at the system/sector level (higher layers) are usually called RRM procedures, while at the user/service layers AMC algorithms. All the optimization scenarios are intersected by the two basic common resources, bandwidth and power, and have as input the feedback information (feedback can contain various forms of information, CSI, error measurements, etc.). The objective function used for each layer is usually called the utility function. Utility functions are also used in cross-layer optimization. They are defined so as to balance efficiency and fairness when allocating resources in systems with heterogeneous services. Consequently, they can be used to optimize radio resource allocation for different applications and to build a bridge between the physical, MAC, and higher layers.

At the center lies the basic element of a packet-based communication system, namely the ‘packet’. The interpretation of a packet herein is that of the smallest portion of communication entity employed by a given system (containing unique information content). It can either be considered correctly received, or discarded at the end of the receiver’s processing
at the link level, thus determining the performance of the link. This approach is compatible with packet-based systems (like WiMax and LTE); where each packet is characterized by the code rate, block size and constellation used. It can also describe uncoded systems where a symbol can be considered as a packet.

When designing AMC algorithms, the packet Link-Level Performance (LLP) prediction under some CSI information is a fundamental requirement for all optimization problems. In some cases of interest (e.g. coded OFDM-based systems), the exact LLP function is difficult to be derived in an analytic form amenable to run-time optimization. This arise the need for a Compact Link-Level Performance Estimation (CLLPE) model that take into consideration the parameterization of the transmitted signal, given the channel and interference conditions. It should be detailed enough to include channel modeling issues such as: the effect of multiple antennas at the transmitter and/or the receiver, the MIMO technique applied (e.g., beam-forming or other spatial multiplexing scheme), and the receiver type. The accuracy of such models is more crucial in lower layers, while rougher approximations can be used at higher layers.

B. Review of AMC techniques

A review of possible adaptation strategies can be found in [21] with extensive bibliography on the subject, as well as a review on the link adaptation research history. A short summary of this review is presented herein. As mentioned before, there are multiple possible combinations of the adaptation criteria, requirements and constraints. For example, in the throughput-oriented strategy, the AMC algorithm aims in providing the highest bit rate (or spectral efficiency) for a required BER and fixed radiated power limit. This was one of the first adaptive transmission schemes proposed by Steele and Webb [22] for single-carrier QAM modulation and narrow-band fading channels. Exploiting the time-variant channel capacity, various concatenated coded schemes with an adaptive coding rate have been investigated in [23]; variable coding rate and power schemes in [24]–[28]; latency and interference aspects with turbo-coded adaptation in [29], [30]. The concepts elaborated for adaptive QAM modulation and coding have been invoked for OFDM QAM in [31], [32]. Adaptive subcarrier selection for OFDM TDMA dynamic links has been investigated in [33]–[35], space-time diversity in [36], [37], and multi-coded systems in [38], [39], as well as in the investigation of the key agents affecting AMC performance [40]–[43]. Interesting proposals for throughput-oriented AMC algorithms can also be found in [44]–[48].

The set of transmission parameters amenable to adaptation is in general large (i.e. constellation size, code rate, Tx power, symbol rate, number of subcarriers, the number of antennas, etc.). Targeting to reasonable complexity is also a decisive factor for the parameters selection. For example in [49] it was shown that by adjusting only the power or only the bit rate the resulting capacity is negligibly smaller than by adjusting both these parameters. In [50] a study on maximizing the spectral efficiency by optimally varying combinations of the transmission rate and radiated power with average power and instantaneous (or average) BER constraints has been presented. There, both continuous-rate adaptation and discrete-rate adaptation are considered. The conclusion has been that the use of only one or two degrees of freedom in adaptation yields spectral efficiency close to the maximum possible that would be obtained by utilizing all degrees of freedom.

In energy-constrained wireless networks a common adjustable parameter is the overall power consumption of a transceiver for a target QoS level. The motivations for this approach are usually: to extend the battery life, to minimize the electromagnetic radiation in populated areas, to reduce the cost in infrastructure-based networks, and to reduce the interference. In this case the transmission parameters should be adapted in order to minimize the power needed for the baseband signal processing and the power of the transmitting antenna. Thus, some trade-offs in choosing the optimization criteria are needed, such as the accuracy of the performance prediction (CLLPE model) within a given limit for the baseband processing. Interesting propositions for power-oriented AMC algorithms and strategies can be found in [24], [51]–[55].

There are interesting proposals in the literature concerning multi-user adaptive schemes. In multi-user systems (for instance, in adaptive FDMA) subcarriers may be adaptively allocated to users in an optimal manner, taking the quality of each user’s channel into account. An example is the algorithm described in [56] which aims at minimizing the overall transmit power by adaptive multi-user subcarrier, bit, and power allocation. In [57] a general link and system performance analysis framework is developed that is used to compare the downlink performance of fully loaded cellular system with different types of link adaptation.

IV. AMC Example Using CLLPE Model

In section III the notion of CLLPE was introduced, as a fundamental requirement for all optimization design strategies. In this section the usefulness of proper CLLPE modeling for AMC design in a coded OFDMA system in a frequency selective environment is demonstrated through an example. This example is chosen because it is compatible with emerging standards like LTE-ADV and WiMax, and the exact LLP function is very difficult to be derived in analytic form amenable to use in run-time optimization.

Finding CLLPE models for coded OFDM systems has been an active area of research and has received considerable attention in the literature [58]–[66]. The main motivation behind these efforts was to use the so-called physical-layer abstraction in order to determine the performance of a given link, and thus to avoid the need for extensive simulation for system level performance assessment. Various systems under development and the corresponding standardization bodies introduced Evaluation Methodologies (EVM) documents which summarize these PHY abstraction methodologies for the respecting systems. For example the EVM Document of IEEE 802.16m [67] provides CLLPE models for coded OFDMA designs that can be used both for system level performance
assessment and AMC algorithmic design. In this case, the role of a CLLPE method is to predict the coded Block Error Rate (BLER) for a given received channel realization (frequency selective) across the OFDM sub-carriers used to transmit the encoded packet. The post-processing SNR values at the input to the FEC decoder are mapped to a single number, called effective SINR SNR\text{eff}, which is then used to accurately relate the system-level SNR onto predefined AWGN link-level curves for determining the resulting BLER. This approach is referred as Effective SNR Mapping (ESM), and based on the adopted model can be categorized in: (a) Mutual Information ESM (MI-ESM) and (b) Exponential ESM (EESM) techniques. The most popular MI-ESM techniques are Received Bit Mutual Information Rate (RBIR) and Mean Mutual Information per Bit (MMIB). The expressions for calculating these methods for the WiMax OFDMA system can be found in [67].

Targeting in the AMC design for Bit-Interleaved Coded OFDM systems, a novel link performance prediction method tagged as “cumulant generating function based” ESM (ESM) is proposed in [66]. Differently from the conventional ESM methods, ESM relies on an accurate evaluation of the Pairwise Error Probability (PEP) figure through the statistical description of the BIC log-likelihood metrics.

In the example presented herein we use the MMIB method as a CLLPE model and the WiMAX-compatible basic parameterization. The channels used are the AWGN (frequency flat) and the frequency selective Pedestrian B Tap Delay Line (TDL) Scenario as described in [67]. Channel coding is based on the WiMax compatible Convolutional Turbo Codes (CTC) as described in [68]. The AWGN link-level BLER performance curves for the 32 CTC modes with varying block sizes (table 524 of [68]) are displayed in Fig. 2. This figure demonstrates the large variety of the available operation modes in the WiMax system and dictates the need for accurate performance prediction in order for a selection mechanism to make the best decision.

If we also take into consideration the various MIMO techniques available in the standard, then the number of possible modes is even higher. When using linear receivers, the overall MIMO channel translates to an equivalent SISO; thus, the problem of ‘optimally’ choosing the best mode is simplified, allowing the use of AMC techniques for an equivalent SISO channel. Thus, the incorporation of linear MIMO receivers does not lead to any fundamental change in the AMC design procedure, other than increasing the allowable Tx-modes. In the case of Maximum-Likelihood (ML) types of receivers the situation is more complicated. Theoretically, MI-based ESM techniques perform the abovementioned translation to equivalent SISO even for ML receivers, enabling the seamless incorporation of MI-based AMC techniques. In practice however, the model loses its “compactness” even in the case of a 2x2 system, making it useful only for performance prediction and not for real-time parameter adaptation.

In Fig. 3 we plot the actual (simulated) performance in the AWGN channel, and the actual and predicted (by using MMIB) performance in the frequency selective channel with and without Bit & Power Loading. The performance without using any bit-power allocation algorithm is depicted as C-BPL (Constant Bit & Power Loading). The MMIB-BPL is a scheme using bit-power loading based on the MMIB ESM technique. The mutual information (MI) of the coded bit depends on the actual constellation mapping of the symbol and the SNR. The MMIB per symbol is a function of the SNR, and is defined as the mean MI of all bits of a symbol; the exact expression can be found in [69]. In the algorithm presented in Fig. 3 an approximation of the MMIB is used via a parameterized sigmoid function [69]. The bit-power allocation is performed by using this approximation along with the iterative solution proposed in [70] for the power loading and Willink’s solution for bit-loading [71], both with proper modifications. Details for the algorithm can be found in [69].

The results are displayed for two selected modes: 4-QAM with CTC rates of 1/2 and 3/4 and one sub-channel (48 sub-carriers) per codeword. In this simulation scenario the target is the performance assessment of a particular mode and the assessment of the mode selection procedure based on a chosen metric. Each SNR point represents the average SNR across each block (sub-channel) used and not the long-term average SNR of the system. The channelization is chosen to be the Partial Usage Sub-Channel allocation (PUSC) in order to address difficult scenarios with high SNR variations across the sub-carriers within a codeword. All channel realizations are normalized in power, thus the performance degradation of the frequency-selective scenario with respect to the AWGN case is only attributed to the SNR variations.

In Fig. 3 we can assess: (a) the degradation due to SNR variations by comparing the ‘Simulation’ curves with the ‘Simulation AWGN’ curves, and (b) the ESM prediction performance by comparing the ‘prediction’ curves with the real ones (‘Simulation’). When the CLLPE model is used to choose the appropriate transmission packet, any deviation in the prediction is a measure for a needed power margin to compensate the modeling uncertainty. Thus, when optimizing the transmission parameters, gain in total power can be achieved not only with the classic ‘shaping gain’ by using BPL techniques, but also from the reduction of the prediction uncertainty.
Classic SNR threshold-based methods lead to worst-case designs, where you always spend excessive power (high power margin) in order to guarantee the target BLER. In the example of Fig. 3 the average BLER prediction based on the MMIB is depicted. For a target BLER of $10^{-2}$, for the case of 1/2 code-rate, the prediction is only 0.7 dB far from the actual performance, while for the case of 3/4 code-rate, the prediction deviation is 1.2 dB. It is important to emphasize that the depicted MMIB prediction is based on the instantaneous channel realization. For example, if the channel is flat, the performance prediction follows accurately the AWGN curve. The difference at the estimation of the mean can be interpreted as the average power loss for not having a perfect CLLPE model.

As expected by using MMIB-BPL, we have a performance gain compared with the C-BPL. An interesting outcome of this comparison is that the use of bit-power loading eventually leads to more accurate prediction compared with the C-BPL. In cases where the prediction under-estimates the BLER, we deliberately insert an SNR margin in order to guarantee the QoS. This is very small for the case of MMIB-BPL as depicted in Fig. 3 compared to the classic average SNR threshold approach. The later can be considered as a constant penalty from the flat (AWGN) curve due to channel selectivity.

V. RESEARCH CHALLENGES

This section presents a small summary of open issues for AMC design in multi-modal, multi-parametric emerging standards. A first point of consideration is the selection of the available modes of operation to be used by the AMC algorithm. As the number of modes increase, the optimization algorithm itself can become very complex. An interesting research topic is to find a mechanism to limit the menu of available modes, i.e., to produce some general guidelines for mode selection (as a function of the scenario) without having to necessarily go through an exhaustive analysis.

Since optimization algorithms rely on information provided by measurements and estimated parameters, system robustness analysis is essential. In addition, any algorithmic design with practical interest must incorporate mechanisms that will take into account all possible errors (measurement, feedback, system imperfections, etc.) for the scenarios of interest.

Finally, for systems which involve high-order MIMO schemes using approximate Maximum Likelihood receivers (like WiMAX and LTE-ADV), it is very difficult to derive reduced-complexity link-performance models, even for the simpler case of uncoded performance [72]. Additionally, as per the previous paragraph, parameter estimation errors complicate the problem even more. Thus, developing proper analytic CLLPE models amenable for use in run-time optimization, along with an elaborate trade-off analysis between performance prediction accuracy and complexity is a very challenging part of the overall AMC design procedure.

VI. CONCLUSIONS

The paper presents a literature review on channel capacity and AMC techniques. The AMC framework is also presented, highlighting the need for compact link-level performance modeling in the AMC algorithmic design for multi-parametrical systems. As illustrated in the coded OFDMA example, accurate performance modeling can compensate the performance loss caused by limited parameterization, and for designs targeting guaranteed QoS can significantly reduce the power loss caused by the classic SNR worst-case threshold.

ACKNOWLEDGMENT

This paper is based on the work conducted within WPR.3 of Newcom++ project.

REFERENCES


Considering Microelectronic Trends in Advanced Wireless System Design

Dominique Noguet, Guido Masera, Venkatesh Ramakrishnan, Marc Belleville, Dominique Morche, and Gerd Asheid

Abstract—Wireless communication system design has been a booming topic since the shift into the digital era in the 1990s. In the same period of time, microelectronic technologies have reached new paradigm points as they were going deeper into the sub-micron area. This paper gives an overview of these emerging constraints and enablers, looking through the specific angle of how much this may impact future wireless system design. To this end, the paper analyzes the major requirements from modern digital communication systems, the way it is foreseen to evolve, and how it can be mapped onto the microelectronic roadmap.

I. INTRODUCTION

RECENT advances in digital wireless communications introduced the use of complex and computational intensive algorithms. This is particularly true as far as the PHYsical and Medium Access Control (MAC) layers are concerned. Indeed, a general trend of these digital communication systems is to improve as much as possible the use of the spectrum resource, which is a scarce and expensive commodity. Efficiency means in this case spectrum efficiency (bit/Hz/s) but also coverage, coexistence, and quality of service provision. To that end, many techniques have been proposed over the last decade mainly, such as new modulation schemes (e.g. WCDMA and OFDM), space-time coding (or in a broader sense Multiple-Input Multiple-Output (MIMO)) techniques, channel coding, etc, that have pushed performance close to theoretical capacity limit [1]. This trend has put hardware designers under pressure; as they have to tackle these highly demanding schemes, while coping with power consumption issues and limited evolution of the silicon technology. Considering both the International Technology Roadmap for Semiconductors (ITRS) [2] and the evolution of wireless communication standards, is indeed a good way to understand that the evolution of the wireless world cannot be caught up by the Moores law alone and that new architectural concepts have to be found to fill the gap. This has moved the centre of attention to the exploitation of parallelism and, unavoidably, opened new questions about how to exploit the different levels of parallelism and how to develop consistent interconnection systems between the processing elements. These open questions are at the core of Multi-Processor System-on-Chip (MP-SoC) research. These systems try to exploit independent-tasks (or functional parallelism) by mapping them to a large number of processors, interconnected via a proper communication structure (e.g., Networks-on-Chip, (NoC)). Considering power consumption leads to even more stringent requirements since battery technology moves yet at a slower pace. As it was stressed above, wireless technology moves fast and more and more standards are to be considered at design time and used/maintained over their lifetime. This makes flexibility a must for current transceiver design. Over the past few years, chipset and equipment manufacturers have adopted a platform approach for the design of a new release to enable to consider the evolution between standards in an incremental efficient fashion in which a new chipset is considered as an evolution of its predecessor rather than a brand new design. However, such a methodology, though using a flexible approach at the design stage, does not necessarily lead to a flexible instantiation eventually. Yet, another approach consists in considering that a transceiver needs to handle flexibility in operation. This interest has been increasing over the past years and is referred to as Software Defined Radio (SDR) [3]. In fact, two levels of flexibility can be considered. The first one captures the ability of a transceiver to support a variety of different modes within a given standard. This is used for instance in adaptive modulation and coding schemes. The second one relates to the fact that a modern transceiver has to handle different standards that are switched from one another depending on availability and user needs. However, current solutions still exhibit low performance either in terms of flexibility or in terms of power consumption. This paper is an attempt to analyse trends in microelectronics in order to better understand the emerging constraints and enablers future wireless system designers need to consider. In Section II, the evolution of microelectronics is depicted, showing the constraints appearing in deep sub-micron CMOS technology as well as potential enablers from technologies that are aside the traditional Moores law track. Then, the specific requirements coming from the communication system design is highlighted in Section III, with a focus on key parameters that are directly impacted by the underlying microelectronic technology. A specific requirement that comes from the profusion of different standards is the thirst for more flexibility. This has to be addressed both at the digital and the RF level, though with specific solution in each case. This is the specific scope of Sections IV and V respectively, again considering how this can be helped by emerging technological solutions.

Dominique Noguet, Marc Belleville and Dominique Morche are with CEATEL MINATEC, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France (e-mail: <firstname.lastname>@cea.fr)

Guido Masera is with Dipartimento di Elettronica Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino - Italy (e-mail: guido.masera@polito.it)

Venkatesh Ramakrishnan and Gerd Asheid are with Institute for Integrated Signal Processing Systems, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany (e-mail: <firstname.lastname>@iss.rwth-aachen.de)
II. TRENDS IN SILICON TECHNOLOGIES

Wireless technology has benefited from the advances the silicon technology has offered in the 1990s and 2000s. The whole telecommunication mutation from the analog domain to the digital realm has in fact been made possible by the miniaturisation of transistors, leading to higher density though with low power efficient ICs. In turn, this move to digital communication has created the boom of the digital ICT at all layers, from broadband communications to multimedia services. With this in mind, it would not make sense to foresee what telecommunication will offer in the future without considering the trends in silicon technology research and industry. The key factor behind the digital revolution has been the CMOS technology down-scaling following the so-called Moores Law which coined in the early 70s that the number of transistors would double every year. Although this rule has been validated over the 40 last years by the silicon industry and by the ITRS [2], it is agreed that we are coming to a new era where this rule is no longer valid. Several reasons can be identified:

- The physics of silicon introduces side effects in deep sub-micron technology,
- Predictability of transistors behaviour is getting less accurate, leading to lower yield or less optimal usage of silicon,
- Power density is going to levels beyond what cooling can offer,
- Static power increases which makes no longer valid the assumption that the overall power consumption decreases as transistors shrink,
- Investments needed for new deep sub-micron CMOS are being so huge that only less than a handful of application justifies it.

For all these reasons, it is likely that we are on the verge of significant changes in the silicon capability roadmap, which make the analysis of future trends useful. Indeed, it is foreseen that the roadmap has to move from a pure down-scaling to new functionalities and combined technology vs. system innovation in order to manage future power, variability and complexity issues. However, there is no accepted candidate today to replace CMOS devices considering the four essential metrics needed for successful applications: dimension (scalability), switching speed, energy consumption and throughput [4], [5]. Moreover, when other metrics such as reliability, designability, and mixed-signal capability are added, the dominance of CMOS is even more obvious. It is then realistic to think that other micro or nano-technologies should be seen as future add-ons to CMOS and not as a substitute for it [6]. This transition between the “business as usual” era and the entry to the post 2015 period where new alternative or complementary solutions need to be found is depicted in Fig. 1.

Bearing in mind this disruptive future and rather than extending the technology evaluation proposed by the ITRS,

1 2015 is generically considered as the end of CMOS scaling because it has been shown that channel length will reach dimensions where the MOS device principle no longer operates.
3 Source: IMEC, 2006.

Fig. 1. Technology nodes and predicted end of CMOS down-scaling with possible alternatives (nanotubes and nanowires) for post-2015 nanotechnology

technology analysis carried out in Europe by EUREKA Medea experts [6] suggests considering 3 major paradigms:

- **More Moore**: corresponding to ultimate CMOS scaling
- **More than Moore**: corresponding to the use of heterogeneous technologies such as Micro Electro Mechanical Systems (MEMS) or Micro-Opto-Electro Mechanical Systems (MOEMS)
- **Beyond CMOS**: corresponding to nanotechnology alternatives to CMOS.

A. More Moore

This ultimate scaling of CMOS will be essential to supply the massive computing power and communication capability needed for the realisation of European Ambient Intelligence (AmI) applications at an affordable cost and a power efficiency exceeding 200 GOPS/Watt for programmable and/or reconfigurable architectures [7]. However, reaching this ultimate CMOS node at the deca-nanometre level around the year 2015 will require addressing cumulative interrelated challenges surveyed in [8], [9].

In the process technology domain, major challenges are: the massive introduction of new materials, the introduction of new device architectures, the move to Extreme Ultra-Violet (EUV) litho or nano-imprint lithography, the increase of random device and interconnect variability especially in memories, the reach of limit of Cu interconnects (e-migration, cross-talk, etc.), the conflict between dynamic and static power density.
In the design domain, the key challenges are: the fact that ultimately non recurring engineering (NRE) cost may reach $1 \text{B} \Omega$/platform if no drastic changes in design technology occur, due to increased hardware-software interaction on multicore platforms. Perhaps even more relevant is the fact that process technology challenges directly impact design challenges. One example that is relevant for handheld terminals is that static power will become prominent in the energy consumption bill. Unfortunately, it is not expected that battery power density will evolve at the same pace (see Fig. 3).

Random variability will impact parametric yield and will require novel ways to avoid corner-based design to cope with device uncertainty, and amenable to design automation. This will require the development of self-healing, defect- and error-tolerant, yet testable design based on low-cost on-chip adaptive control systems.

Reliable local and global on-chip communication in 22 nm or smaller technology will be a much more limiting factor than transistor scaling and will require, besides the investigation of optical, wireless or CNT-based technologies, investigation of architectural solutions such as tile-base Globally Asynchronous Locally Synchronous (GALS) architectures exploiting Networks-on-Chip and MP-SOC. 3-D integration and System in Package (SiP) must also be studied as strong contenders to ultimate scaling for true system design, which is finally the ultimate goal of electronics.

Analog and RF design will have to cope with ultimate digital scaling and further sub-1 Volt scaling. This will require extreme creativity in analog and RF system design by compensating analog deficiencies by digital techniques.

Alternatives to bulk CMOS shall also be considered to overcome the shortcomings when scaling down to deep submicron. Silicon On Insulator (SOI) technologies are foreseen as relevant with this regard, as it could lead to a better tradeoff between active and static power leakage. A significant active power reduction can be achieved by using SOI devices. Indeed, SOI devices are well known to be able to achieve the same performances as Bulk device, but with a lower power supply. This is achieved thanks to lower parasitic capacitances (thick buried oxide for electrical isolation instead of junctions) and thanks to lower threshold voltage in dynamic mode4. Active power reduction up to 50% can then be achieved with SOI. DC leakage can also be controlled [10].

B. More than Moore

The “More than Moore” approach intends to address parallel routes to classical CMOS by tackling applications for which CMOS is not optimal. These applications can be classified in three major categories: interfacing to the real world, enhancing electronics with non-pure electrical devices, embedding power sources into electronics [6].

In the field of IC design for advanced 3G standards, the More than Moore class is expected to bring significant breakthroughs in RF front end design. Indeed, new complex signal modulations (e.g. OFDM) require very linear RF components in order to limit distortion and to ensure high signal throughput. Cellular phones can utilise up to seven different wireless standards or bands, including DCS, PCS, GSM, EDGE, CDMA, WCDMA, GPS and Wi-Fi, and each standard has its own unique characteristics and constraints. Additionally, next generation phones cannot be significantly larger than todays phones and they will need to have similar talk and standby battery lifetimes. Today, a large proportion of the components in a mobile phone are space consuming “passive” elements such as inductors, variable capacitors and filter devices. Integrated passive elements and RF MEMS/NEMS have been proposed to help solve these problems. High-quality passive elements are available through SiP technologies. Besides, nano-materials are expected to strongly improve the achievable capacitance per unit area value for capacitors. One key remaining issue though, is whether the extra cost of these non-standard technologies can be justified by the benefits they bring. Thus, whenever heterogeneous technology is considered, the trade-off between performance and cost must be analysed.

4In case of Partially Depleted SOI devices.
The “Beyond CMOS” paradigm intends to identify technologies that could replace CMOS, either in a disruptive or evolutionary way after CMOS will reach its ultimate limits. The ITRS Emerging Research Devices (ITRS-ERD) proposes criteria to evaluate the potential of emerging research devices and circuits with respect to future applications. The analysis presented in the ITRS-ERD document [2] is based on defining a set of criteria for logic and another set of criteria for memories, and applying them to potential technologies. These criteria are as follows:

- **For Logic**: scalability, performance, energy dissipation, gain, operational reliability, operating temperature, CMOS technological and architectural compatibility.
- **For Memories**: scalability, performance, energy dissipation, OFF/ON ratio, operational reliability, operating temperature, CMOS technological and architectural compatibility.

Nanotechnologies falling into this category correspond to building blocks such as: Atom scale technologies, Spin electronics, Molecular electronics, Ferromagnetic devices, Nanoelectromechanical systems, Organic/plastic electronics, Bio-sensors. Because these new devices have behaviours that sometimes differ significantly to classical transistors, “electronics” using these functions needs to be invented as well. From an architectural viewpoint, these technologies derive into: bio-inspired electronics, nanomechanical computing, quantum computing. Stating this, it is obvious that a description of the research challenges related to Beyond CMOS is far too broad to be surveyed in this paper. What can be kept in mind is that Beyond CMOS is extremely multi-disciplinary with extensions at all levels from building blocks to system usages.

III. PROCESSING NEEDS IN WIRELESS COMMUNICATION

Bearing in mind the evolution of silicon technology, designers also have to consider the evolution of the require-ments from advanced wireless systems. Along their path towards next generation broadband wireless access, different standardization bodies (e.g. 3GPP, 3GPP2, IEEE, ETSI-DVB, etc.) have been introducing new standards that enhance their legacy radio access technologies. Examples of recent releases are: 3GPP Release 7 (HSPA+), Release 8 (LTE), Release 9 (LTE-Advanced), 3GPP2 (UMB), IEEE WLAN 802.11 (n, vht), IEEE WMAN 802.16 (d, e, m), ETSI DVB (T2, H, SH, NGH). Emerging and future wireless communication systems are characterized by a clear and steady convergence on both services and technologies. From the service perspective, operators are striving to offer to the users a wide spectrum of rich multimedia services including both interactive and broadcasting, which raise the need to embed complementary technologies (e.g. for uni-cast, multi-cast, and broadcast trans-missions) into the future generations of radio access systems.

This context of coexistence and convergence is driving demand for flexible and future-proof hardware architectures offering substantial cost and power savings. Manufacturers have already started activities towards the provision of multi-mode handsets featuring the advancements of the recent radio access technologies (e.g. 3GPP LTE, WiMAX IEEE 802.16m, DVB-T2/H). However a large gap is growing in the field of flexible radio between advances in communication algorithms, methods, system architectures on one side, and efficient im-plementation platforms on the other. Despite the large amount of available results in the system level technologies related to multi-mode, multi-standard interoperability and smart use of available radio resources (adaptive coding and modulation, cross layer optimization, ...), a very limited number of hardware solutions have been proposed to really support this flexibility and convergence by means of power efficient, low cost reconfigurable platforms. In most cases, chipset vendors offer different solutions for each combination of standards and applications to be supported.

To enable a single modem to service multiple different wire-less systems, highly flexible solutions are needed. In practice, currently implemented flexible hardware modems are focused on the receiver segment placed between the RF front end and the channel decoder. In this part of the modem, several digital signal processing algorithms, such as equalization, interference cancellation multipath correlation (rake receiver), synchronization, quadrature amplitude mapping/demapping and Fast Fourier Transform (FFT) can be run on vector processors, which allow for Giga-Operations Per Second (GOPS) rates resorting to very high level of parallelism. However, other functional components of a modern modem (such as channel decoding) are not efficiently supported by vector processors and alternatives to software programmable architectures are not considered solid solutions: cost of hardwired dedicated building blocks becomes rapidly unacceptable with the number of standards to be supported, while reconfigurable hardware, such as Field Programmable Gate Arrays (FPGAs), are too expensive in terms of silicon area and standby energy con-sumption (which is due to leakage current, proportional to area).

The design of high throughput software programmable architectures is currently the largely prevailing development
Beside this research effort gap, known platforms for digital base-band processing show serious lacks of capabilities with respect to the radio flexibility that is currently studied at the system level and expected by the market. In particular presently available platforms suffer from two main limitations:

- Partial flexibility: since large difference is recognizable between the processing functionalities that are expected to be supported in a flexible receiver, and the actual level of flexibility that is achieved in hardware. As an example, advanced channel decoders are usually not included in the whole receiver as reconfigurable elements, but they are supported by means of separate components, designed and optimized to execute one specific decoding algorithm.

- Expensive flexibility: since flexibility comes at a very high cost in terms of occupied area and dissipated energy and required reconfiguration time; on the other hand, flexible platforms are requested to provide better overall die area and power figures than receivers designed by simply allocating multiple function-specific components. Moreover, simple, fast and energy efficient reconfiguration procedures are of primal importance to enable true flexibility, as awaited in cognitive and opportunistic radio systems.

A. The need for highly demanding building blocks

A rather reduced number of computationally intensive functionalities is associated with the key enablers commonly adopted by next generation radio access standards. A list of these technologies is presented in Fig. 5. The table on the left shows key characteristics of Forward Error Correction (FEC) technologies adopted in several standards that have been introduced during the last 15 years in the domain of wireless communications and digital broadcasting. Particularly Fig. 5 contains for each specified FEC the maximum data throughput and the processing complexity, expressed in GOPS. The plot on the right side clearly indicates that both throughput and complexity tend to increase exponentially with time. In more details, data throughput doubles within 15 months and this trend is almost in agreement with the evolution trend of performance in semiconductor industry, at least until Moore's law remains valid (see section 2). However, Fig. 5 shows that the FEC complexity trend is faster, as the required GOPS doubles every 12 months.

The growing gap between silicon performance and FEC complexity trend implies that the efficient implementation of emerging and future error correcting techniques will not be guaranteed by the progress in the semiconductor industry, but will continue to impose application specific optimizations at the confluence of algorithm and architecture. Particularly in the implementation of computationally intensive base-band processing tasks, this joint effort at the algorithm and architecture levels can be targeted towards different optimization objectives, such as area occupation, throughput, power dissipation, and flexibility.

As for the area and throughput objectives, several cases can be cited to show how efficient implementations often come from joint design efforts spent at the algorithm and architecture levels. For example, although the original Low Density Parity Check codes (LDPC) decoding algorithm requires rather complex processing at the check nodes, all decoders implemented in the last few years resort to the min-sum approximation, which saves a significant portion of complexity with a marginal performance loss. Another relevant example is found in the MIMO detection domain, where several suboptimal implementations have been proposed in the last few years as alternatives to sphere decoding. These solutions (e.g. k-best and LORD) exhibit close to Maximum Likelihood performance, simpler architecture, and deterministic detection delay at the same time. Additionally, the algorithm-
architecture interdependency in a specific application domain can be leveraged to improve energy efficiency.

Joint algorithm and architecture optimization can be also seen as a method to achieve flexibility through the development of unified processing algorithms that enable increased sharing of hardware resources. This unifying approach combines both algorithm and architecture alternatives into a joined optimization effort towards efficient flexibility in radio communications. The idea is to develop innovative processing algorithms, where limited and controlled performance degradation is accepted and exchanged for flexibility in the hardware implementation.

B. The need for power efficient design

Several IC implementations of turbo and LDPC decoders have been published in IEEE ISSCC International Solid State Circuits Conference (ISSCC) and in the Journal on Solid State Circuits (JSSC) series in the past few years. Looking at the characteristics that have been measured for those components, one can easily see that the power dissipation trend-line remains fairly constant across the last ten years and typically included in the range between a few tens and a few hundreds of mW. This general tendency is somehow surprising, as throughput and processing complexity have been increasing along the same period of time. The explanation of the observed trend on power dissipation comes from CMOS process down-scaling (particularly scaling of gate area and parasitic capacitance) that has substantially balanced the increase in computational effort. This appears as a very bright and encouraging conclusion. However power dissipation is expected to become a very critical issue in future developments for several reasons.

First of all, the static power dissipation has been usually neglected so far; however it will soon become comparable to the dynamic one in next generations of CMOS process technology (see section 2). Therefore, with a fixed power budget assigned to base-band components, the dynamic power consumption will need to be limited to a lower bound than in today implementations.

Secondly, some base-band functions will increase dramatically the need of processing energy. The most relevant example is probably given by joint MIMO detection and channel decoding. The concatenation of soft output MIMO detection and iterative error correction algorithms will create high complexity receivers where the FEC processing is organized around two nested feedback loops: an inner loop, associated with turbo or LDPC decoding iterations, and an outer loop, with exchanged soft information between MIMO detector and inner channel decoder. This arrangement will significantly increase the global complexity and affect both the required processing speed and dissipated power. For example, if a 300 mW turbo decoder is used as the inner unit in a concatenated system with 3 iterations of the outer loop, the power consumption of the turbo decoder will increase by a factor of 3.

A third reason comes from the increasing levels of flexibility that will be incorporated in next generations of channel decoders. Flexibility is ineluctably associated to a cost in terms of consumed energy, which tends to reduce the power efficiency.

The rise of power consumption combined with the wished reduction in size of handset devices causes temperatures to increase because the transfer of heat is proportional to the surface area. Increased temperatures have two effects. The first is that the temperature of the casing of the device can go up such that it becomes too hot to handle. The second effect is that higher temperatures make the electronic components unreliable and more likely to fail. In [11] it is envisaged that a dramatic increase of energy consumption of 4G mobile device will make active cooling a necessity, which is not attractive for users and manufactures. The performance of active cooling in a mobile devices is investigated in [12]. From the mobile manufacturers’ perspective the energy consumption problem is critical, not only technically but also taking into account the market expectations from a newly introduced technology. This is in fact becoming a key concern: there exists a continuously growing gap between the energy of emerging radio systems and what can be achieved by:

- Battery technology evolution,
- Scaling and circuit design progress,
- System level architecture progress,
- Thermal and cooling techniques.

Considering this power consumption (and dissipation issue), designers are more and more considering power consumption as a key figure of merit. A simple model can help clarify this issue and at the same time suggests the main research direction to combat loss of efficiency in flexible architectures. In channel decoders, power efficiency is defined as the ratio between the number of decoded bits per second and the corresponding dissipated power. It is usually measured in Mbps/mW. The numerator of the ratio can be written as $N_b \times f_{ck}$, where $N_b$ is the number of bits decoded per clock cycle and $f_{ck}$ is the clock frequency; neglecting the static contribution, the denominator can be expressed as $A \times C_{sw} \times V_{dd} \times f_{ck}$, where $A$ is the total occupied area, $C_{sw}$ is the average switched capacitance per unit area and $V_{dd}$ is the supply voltage. Defining $A_b = A/N_b$ as the average area required to process one bit, the power efficiency can then be formulated as $\eta = 1/(A_b \times C_{sw} \times V_{dd}^2)$.

This simple model clearly shows that the power efficiency only depends on three parameters: two of them, $C_{sw}$ and $V_{dd}$ are tight to the evolution of silicon technology, while $A_b$ is largely dependent on how the required processing functions are implemented. When two or more functions are mapped to a unique architecture capable of flexibly support all of them, this necessarily leads to the allocation of additional components that are used to handle the switching between two functions and not for the execution of the function itself. This results into an increased average area per decoded bit and impairs the efficiency. Therefore the search for efficient flexibility is the search for architecture solutions that minimize $A_b$.

C. The need for flexibility

Although throughput and area have been the dominant metrics driving the optimization of digital building blocks, recently, the need for flexible systems able to support different operative modes, or even different standards, has changed the perspective. In particular, the SDR paradigm made flexibility a fundamental property of future receivers, which will be requested to support a wide range of heterogeneous standards.

Run-time flexibility in a receiver is a very ambitious and innovative task that shall provide support to multiple versions of a specific functionality, each one characterized by a different trade-off between communication performance and energy (or throughput) efficiency. The fundamental purpose here is dynamically enabling the change between one version and another of the considered functionality, in response to energy constraints and user needs. This type of versatile platform will therefore give support to complex power management algorithms and optimal allocation of the spectrum resources. Although the concepts of adaptive and cross-layer optimization in mobile terminals are not new, the design of an implementation platform supporting these concepts is still an open problem and a very challenging research topic. Two key problems can be seen: the required level of flexibility is higher than in classical multi-standard architectures, and constraints on the reconfiguration latency are expected to be stricter.

Flexible algorithms and architectures must be developed to enable the support of energy management techniques involving all functionalities of the digital base-band processing chain. Specific optimization metrics and methods need to be introduced to drive algorithm and architecture design. Proper methods must also be developed for estimating the operative conditions and algorithms for realizing the energy management.

A relevant application example for the mentioned flexibility target is given by next generation wireless systems that use multiple antennas to deliver very high data rate services. In such systems, a feedback loop from MIMO detector and outer channel decoder enables iterative “Turbo-MIMO” processing: performance very close to a posteriori probability detection has been achieved with different detection techniques. While optimum error rate performance is obtained with soft-output maximum likelihood detection, linear and successive interference cancellation (SIC) algorithms are interesting alternative solutions, with an implementation complexity lower than the sphere decoding. Moreover, for low code rates or low modulation order, plain linear detection performs within 2dB of the performance bound; on the other hand advanced detection strategies (sphere decoding) is convenient for high code rate and higher order modulation [13]. Thus different performance-complexity-energy trade-offs are covered by a set of heterogeneous algorithms and a proper flexible platform is required to dynamically exploit the energy minimization opportunities offered by these trade-offs.

Current approaches to multi-standard functionality pragmatically aim at implementing a “just enough flexibility”, by supporting codes and throughput requirements specified in some of the current standards. A rather small number of multi-standard decoders have been implemented so far. They can be classified in three categories.

1) PIntra-family flexibility, which support multiple modes belonging to the same functionality. As an example, one could design a turbo code decoder able to operate over several turbo codes, specified in different standards, such as UMTS, WiMAX and WiFi. The most common implementation approach for this category is hardware parameterized functions: the processing architecture is organized around a number of storage and computation units that are structured based on a number of parameters, such as block size and code rate.

2) Inter-family flexibility, which is capable of a wider flexibility, as they must process functions belonging to different and in some cases heterogeneous families (e.g. to stay with the FEC example, turbo and LDPC codes could be an example) specified in two or multiple standards. In this case, reusable hardware resources can be identified and shared among supported decoding algorithms, with the final objective to save area with respect to the straightforward allocation of several independently designed decoders (“Velcro approach”).

3) Full flexibility, which supports high throughput implementation of a wide range of heterogeneous functions, not necessarily limited to the ones that are today specified in a standard. As an example, a fully flexible turbo code decoder should be able to support any interleaving law. The additional difficulty of this approach derives from the fact that parallel collision-free decoding architectures are heavily based on the specific features of the code family to be decoded; as a consequence, these architectures can hardly be exploited when multiple different codes must be supported. Common operator technique is another approach belonging to this class [14].

IV. SOFTWARE DEFINED RADIO APPROACH: CHALLENGES AND OPPORTUNITIES

Flexibility requirements, when coupled with low-cost and less time-to-market constraints, make the development of a mobile device highly complicated and challenging. SDRs, with cognitive capabilities, are getting prominence as potential candidates to meet the future requirements of mobile wireless devices. Compared to the pragmatic design approach for flexibility mentioned above, the SDR approach aims at providing
a comprehensive design framework encompassing platforms, architectures, software, methodology and design tools.

A. Current Solutions

Current solutions for SDRs are component based and model driven, where a Platform Independent Model (PIM) of a waveform is constructed as an assembly of components, e.g. [15], from the specification document, e.g. [16]. Each component represents a part of functionality in a whole waveform. From a PIM model, a Platform Specific Model (PSM), which denotes the implementation of a waveform, is obtained with or without using a library. Libraries internally developed or from third party vendors providing efficient implementations of few, sometimes even all, components of a waveform can improve overall system efficiency and drastically decrease development time. For example, Texas Instruments (TI) provides efficient implementations for implementing the WiMax waveform [17].

Some other approaches for developing SDRs are based on Software Communications Architecture (SCA) adopted by JTRS [18]. SCA is predominantly General Purpose Processor (GPP) based and uses CORBA as middleware abstracting the underlying hardware. This creates an operating environment that enables to develop applications independent of hardware, and methods for loading new applications, configuration and control.

B. Key Problems

Though each of the above solutions improves the development of SDR in one way or the other, there are associated issues often leading to situations where one solution does not fulfil all requirements. A serious drawback in the libraries that are available today on the market is that they are specific to one waveform or to one hardware platform. For example, the library in [17] is specific to the WiMax waveform targeting a specific Processing Element (PE), namely the TMS320TC6482 DSP. Moreover, some of the libraries are proprietary in nature; the details on the components and their interfaces are not known. Even though library based approaches have the potential to increase efficiency and portability, the lack of standardization decreases reuse of implementations drastically.

Similarly, overhead caused for supporting SCA is a key deterrent for its usage, particularly in physical layer processing, due to the existence of hard constraints, e.g. latency [19]. Though abstraction of the underlying hardware platform makes mapping of a waveform description onto a hardware platform easy, it completely blocks the opportunity to exploit the architectural capabilities of a hardware platform. Hence, an optimum mapping is not possible. Mapping should consider the requirements of a waveform, e.g. processing complexity, available resources in a hardware platform, e.g. memory, and constraints of a waveform, e.g. throughput, when optimizing with respect to design requirements like energy efficiency. Therefore, constraint aware mapping is a key for improving the overall efficiency of the complete system.

The efficiency of a waveform implementation is a pivotal factor for overall footprint and energy efficiency. Investigations done on implementation efficiency indicate optimization limits depending on the implementation type [20]. For example, implementing in assembly is more efficient than C-code because assembly code can better exploit the architecture of a PE. A GPP offers high flexibility, but requires more energy per decoded bit than, e.g. a Digital Signal Processor (DSP). Therefore, entirely GPP and C-based SDR solutions are not suitable for battery operated devices due to low implementation efficiency.

Merely increasing parallelism in order to increase computation power, without considering efficiency will lead to high area and energy consumption. Therefore, in SDR systems where future requirements of computational performance will be in the order of tens to thousands of GOPS, techniques like massive pipelining, increasing the number of GPP cores or speeding up the clock are not satisfactory due to energy efficiency reasons. HW platforms for SDRs will, most likely, be heterogeneous in nature with programmable PEs like Application Specific Instruction-set Processors (ASIPs), DSPs, GPPs and Application Specific Integrated Circuits (ASICs) or even physically optimized ICs. Design requirements of a SDR system, including flexibility and efficiency, will determine the type and number of PEs. For example, physically optimized ICs provide very high performance and power/energy efficiency; however they offer least flexibility whereas GPPs provides full flexibility at the cost of very low energy efficiency.

Numerous issues in waveform development for SDRs are due to the “specification-to-implementation” problem. In general, waveform specification is in the form of a textual document with details on different modes, constraints and critical loops that have to be met by a waveform implementation. Textual documents provide redundant information, which is sometimes verbose and sometimes terse. Therefore, creating a PIM of a waveform incorporating all the features like latency and deriving a PSM model, meeting key requirements like throughput from it, is a cumbersome task, often error prone. Therefore, design, development, integration and testing of waveforms have become highly complex and time consuming.

C. Challenges

The paradigm of SDR poses new challenges or makes current design challenges more stringent. The most relevant ones are:

- **Portability**, which can be defined as the inverse of porting effort, represents the ease with which one waveform can be moved to another hardware platform [20]. Portability requires a platform independent waveform description.
- **Efficiency** with respect to area and energy is essential in order to decrease the power/energy consumption and extend the battery life. However, this requires high efficiency in waveform implementation.
- **Interoperability** denotes the ability that a waveform implemented on two different hardware platforms interoperates with each other.
- **Loadability** illustrates the ease with which a waveform can be loaded, over-the-air, into a hardware platform.
programmed, configured and run. Loadability can be increased by well defined and known interfaces in waveform implementation.

- **Trade-offs** between flexibility and efficiency becomes challenging in the wake of their contradictory nature. This makes heterogeneous multi-processor system-on-chips (MPSoCs), an inevitable candidate as the hardware platform for implementing a waveform.

- **Cross layer design** and optimization techniques are getting popular, if not mandatory, in order to cope with the increasing need for spectrum and energy efficiency. This leads to very tight dependencies, interactions between physical and MAC, higher layers that have cognition, requiring flexibility in implementation and algorithms.

Most of the challenges in SDR arise due to the contradictory requirements of flexibility, performance and efficiency. Heterogeneous MPSoCs with specialized PEs can pave the way to solve the dilemma of contradicting demands of high computational performance at one hand and energy efficiency on the other. However, designing such a system is a challenging task. Tools are required for the development of the dedicated PEs as well as of the whole SoC. High speed simulation is necessary in order to support design space exploration and verification at an early phase.

Still, the complexity of modern flexible implementation structures would hardly be manageable and their development is a tedious and error-prone task. What is needed is a description method that can lead to a (semi-) automatic generation of a waveform implementation directly from the specification. Therefore, a methodology is required, that raises the abstraction level of receiver design to make it manageable.

**D. Opportunities**

As mentioned earlier, direct implementation on a low abstraction level is not well suited for an efficient portable waveform implementation. Raising the abstraction level leads to library based approaches, where efficient implementations of basic components are available and can be assembled to implement the complete transceiver. Also, a library based approach enables efficient utilization of heterogeneous MPSoCs.

1) **Design Principles**: There are several key design principles that must be considered while building a library that can pave way for Waveform Description Language (WDL) based SDR development. They are:

- Hardware architectures that offer full flexibility, e.g. GPPs, are not efficient and are costly in terms of area and energy consumption. Therefore, application specific optimization is needed in order to increase both energy efficiency and computation performance. If limited flexibility can be offered, such architectures can still be tuned for different requirements.

- Algorithms that might work efficiently for one scenario might not be efficient for another, e.g. sophisticated and complex algorithms might be needed in a bad channel while simple algorithms might be sufficient in a good channel. This creates the need for analyzing algorithms that are scenario-specific and to identify the common kernels in these algorithms (“Nuclei”) to maximize reuse.

- Building a library that is based only on functionalities in waveforms limits reuse of the library. For example, if one of the components in a library is a modulator of a particular scheme, a different scheme in another waveform renders it useless. Therefore, emphasis should not be on the functionalities but on the algorithms that are used for implementing such functionalities. This not only increases reusability, but also provides algorithmic flexibility.

- Flexibility in implementing a waveform can be provided, even in a fixed hardware platform, through different implementation algorithms and configuration parameters like implementation-method, input data-width, scaling, etc. However, a PE in a hardware platform should have architectural capabilities to support different implementation algorithms efficiently.

- Providing easy programmability for complex systems like SDR is essential to exploit efficiently the hardware resources. A programming model that can bridge the gaps between waveform, hardware platform and mapping is needed. This model should allow a designer to utilize the flexibility present in a hardware platform in order to increase the implementation flexibility.

- Due to the presence of a number of layers with very high interaction between them in typical waveforms, it is essential to treat SDR development as a joint optimization problem.

- In spite of advances in standardization due to bodies like SDR Forum [21], NGMN alliance [22], JTRS program [18], etc., lack of complete and unified standardization is preventing huge advances in SDR technology. Due to this, reusability of other solutions, participation of different vendors is limited, indirectly leading to increase in development costs. This also prevents co-operations and sharing knowledge between academia and industry on the one hand and between military and civil domains on the other hand.

Due to the strong dependencies between algorithms, hardware architecture and tools, it is necessary to investigate these aspects jointly in order to identify an efficient SDR development methodology. For example, a detailed algorithm analysis can drive the component identification and implementation, which then feeds back the analysis results which may cause revision of the algorithm itself, making it algorithm architecture co-design. Furthermore, the real implementation of Nuclei has the potential to deliver important information on the interfaces and parameters which are required for tools exploiting the spatial and temporal mapping of a waveform description. Therefore, joint results achieved by working together in the three domains listed before are needed, making SDR development algorithm, architecture and tools co-design.

2) **Algorithms**: Since SDRs have to offer flexibility, it is efficient, if not necessary, to exploit the tradeoffs between complexity and error rate performance in different algorithms. For example, in a spatially multiplexed MIMO signal, though exhaustive search delivers the minimum error rate, the enor-
mous computational complexity is a heavy burden for the base band receiver. On the other hand, low-complexity algorithms such as zero-forcing detection can operate in a limited SNR range only. Therefore, it is essential to analyze such algorithms jointly along with their tradeoffs.

Algorithms that are used for implementing different functionalities can have common computation and communication patterns. This commonality can be exploited by identifying such common kernels that are also computation intensive. Such algorithms might be used in different applications. If the granularity of such kernels is optimum, i.e. not coarse grained as a complete channel decoder nor as fine-grained as an adder, it can enhance reusability and can enable the availability of optimized implementations for such kernels. However, emphasis should be on the implementation-friendly algorithms in order to enable various implementation alternatives, based on different algorithms, without sacrificing efficiency.

3) Tools: Tools must offer a seamless environment for developing SDRs by providing the infrastructure to capture the waveform specification, to do mapping, implementation, integration and verification. Due to a huge number of critical paths involving several components of a waveform, a constraint aware mapping approach is needed. It increases the chances of successful mapping and decreases the number of iterations. However, it complicates not only the tool development but also the identification of the appropriate ways for describing the impacts with respect to constraints.

In order to validate and evaluate the spatial and temporal mapping decisions as well as the performance of the overall system, a system simulation environment in software is needed, which simulates the system behaviour in terms of functionality and timing. The software based system simulation plays a key role in exploration and verification of the spatial and temporal mappings. The information obtained by the simulation can be fed back to the higher layers in order to improve the mapping quality. This can be considered as an iterative approach, which is repeated until a satisfying result is obtained.

4) Architectures: In general, heterogeneous MPSoCs can provide high performance due to parallel processing of tasks and at the same time provide flexibility and efficiency due to the heterogeneous, function-optimized nature of PEs. Among the PEs, ASICs are very attractive candidates for implementing SDR systems, where a fine balance between flexibility through programmability and efficiency through application specific architecture optimization is essential. For example, conventional load store memory architectures may not be able to meet the throughput-latency demands of SDR applications and may become a bottleneck. Therefore, special application specific memory architectures, in addition to other architectural options, are needed to meet these demands. Similarly, due to extremely high throughput and short latency demands in the communication between PEs in an MPSoC, conventional communication schemes like buses are most likely to fail. Instead, the idea of specialized communication architectures, including dedicated links between PEs, even special links that are optimized separately for throughput and latency is gaining more interest.

To summarize, requirements and therefore complexity of SDRs are increasing day-by-day, mainly driven by new applications and services in wireless communication systems. Design and development of a SDR has inherently numerous challenges due to the contradicting nature of flexibility and efficiency requirements. However, this provides tremendous opportunities and calls for a radical change in the way such complicated systems are built. One promising approach, like [23], that has the potential to provide implementation flexibility even in a fixed hardware platform, is the WDL based waveform development using a library of algorithmic kernels. This approach promises not only the participation of vendors by standardization and open interfaces, but also provides algorithmic and implementation flexibility even in a fixed hardware platform.

V. RF TRENDS IN FLEXIBLE RADIO

The digital communication research on multi-standard radio has started based on the assumption of the Software Radio, which extrapolated that RF stages of a radio would be transparent for the baseband processing either thanks to highly flexible RF components or to very high speed converters. Both have shown limitations and further research is needed to achieve highly flexible SDR. In fact, too major approaches emerge for designing a flexible RF. The first one considers very large band RF that can therefore accommodate several systems. This approach suffers from bad sensitivity level though. The second one relies on tuneable components with which parameters can be adapted to match the system requirements. These rules often contradict the guidelines that RF designer are used to consider when defining an RF architecture, which is usually optimised for sensitivity, power consumption, and IC integration.

At the transmitter side, classical approaches usually result in low flexibility architectures sketched in Fig. 7. In such designs, lump elements freeze the circuit performance to given specifications. Thus, this classical circuitry will hardly be adaptable. Multi-standard terminals based on this concept end up with a RF front-end comprising several RF ICs in parallel, also referred to as the Velcro approach. In order to come up with a less costly and bulky approach, new architectures in which the boundary between the analog and digital world has been modified to enable the use of waveform shaping in the digital domain has been proposed.

For instance [24] suggests using a LInear amplifier with Non linear Components (LINC) architecture to efficiently address large Peak to Average Power Ratio (PAPR) OFDM signals. The advantage of this approach is that amplifiers have to handle constant envelop signals despite the non constant envelop nature of the OFDM signal. This leads to better power efficiency and better flexibility, especially when the signals are shaped in the digital domain [25]. Despite its higher flexibility, this architecture still hardly copes with wide band signals and cannot be tuned over a large central frequency range. This is mainly due to the limited flexibility offered by nowadays analog stages.

Similarly, the trend at the receiver side is to limit the number of analog components. The hype for zero-IF or low
IF architectures in the past few years partly came from this trend.

The zero-IF architecture (Fig. 9) has indeed several fundamental advantages over its heterodyne counterpart. The intermediary IF stages are removed and the functions of channel selection and subsequent amplification at a nonzero IF are replaced by low-pass filtering and baseband amplification, for which a monolithic integration is feasible. Although zero-IF exhibit relevant specifications, it suffers from well identified problems such as DC offset, LO leakage and I/Q mismatch (the first being the most prominent one). The low-IF receiver concept has been developed to avoid these drawbacks. Fundamentally the low-IF receiver originates from the conventional heterodyne receiver system. The main difference is that the digitization process is shifted from the baseband part to the IF part. By implementing A/D conversion at this earlier stage, more flexibility at the receiver can be achieved. The concept of low-IF has become even more attractive recently, especially for emerging systems which require higher transceiver flexibility while keeping the terminals’ compact size and energy efficiency. There are several benefits that can be obtained by implementing early conversion, namely: the high degree of programmability at the receiver, and the avoidance of issues associated with analog baseband demodulation, such as I/Q imbalance, DC offset, etc. Despite all these supporting facts, there is still one main obstacle for implementing such architecture: it requires a fast high-bandwidth high-dynamic-range conventional ADC for converting radio signal with sufficient fidelity. Therefore, improving the performance of ADC is crucial to enhance the flexibility of RF receivers. Besides the performance, power consumption of the conversion stages is a matter of concern to integrate such solutions in low power battery operated devices.

Even when limiting the analog part of the transceiver, the use of tuneable filtering is needed, each filter being dedicated to the given bandwidth of the targeted system. RF filtering has always been considered as the bottleneck of the front-end implementation and making it tuneable represents a huge challenge [26]. For instance, Bulk Acoustic Wave (BAW) filters are highly selective band-pass filters that are convenient for a particular application. However, even if BAW filters are tuneable in frequency, this is only limited to a few percents, and tuning control is quite complex to implement in practice. Besides, practical implementations based on Yttrium Iron Garnet (YIG) resonators provide multioctave bandwidths and high quality-factor resonators. However, they consume a significant amount of dc power (1 to 3 W), and their linearity is poor. Moreover they are bulky, expensive and cannot be easily miniaturized for wireless communications. Alternatively, diode varactor-tuned circuits are simple and require little bias current and size, but they have not met the expectations in terms of loss. Solid-state varactors can provide a wide tuning range, but they have loss and linearity problems at microwave frequencies. Therefore, low cost and high performance tuneable solid state resonators is still a myth. Besides solid state solutions, RF MEMS can provide a relevant alternative. Being constructed entirely of low loss metals and dielectrics, these mechanical structures feature inherently low loss properties.

VI. Conclusion

Recent trends in silicon technology and communication system demands exhibit a growing gap between application
needs and what the technology can deliver. A key driver for the telecom industry is the wireless mobile business. Mobility, which relies on battery operated handheld devices, provide stringent requirements on equipments in terms of processing power, power consumption and flexibility. At the same time the battery and silicon technology does not progress at the same pace. The emergence of new standards implementing ever more efficient air interfaces also put stringent constraints on the design time. Thus, the reuse of hardware building blocks and a proper methodology and tools are needed to evaluate hardware performance tradeoffs at the earliest stage. Flexible radio is a promising approach in this regard. However, a unified framework is still to be found to enable the co-design of communication functions and hardware platforms.

VII. ACKNOWLEDGEMENT

Part of this work was performed in the frame of the ICT NEWCOM++ project which is an EU funded Network of Excellence of the 7th European Framework Programme.

REFERENCES


Dominique Noguet (EE’92, MSc’94, PhD’98) joined CEA-LETI in 1998 where he has carried out digital communication hardware architecture research. His research activities cover reconfigurable and flexible radio, and more recently cognitive radio. Since 2001 he had several National and International project management positions in the field of wireless communication. He currently leads flexible radio activities (WPRC) within the European Network of excellence NEWCOM++ and is the technical manager of QoSMOS, a major EU project on cognitive radio. He received a best paper award, and the best PhD award from INPG. He authored or co-authored about 40 papers in peer reviewed journals and conferences. He is currently the head of “digital architectures and validation platforms” ANP group at LETI, where he also leads Cognitive Radio Activities.
Technology as a Need: 
Trends in the Evolving Information Society

Giovanni Emanuele Corazza, Alessandro Vanelli-Coralli, and Raffaella Pedone

Abstract—In this paper, we take a broad view on the Information Society and on the Information and Communication Technologies which constitute its technical and, in a sense, social infrastructure. We discuss the fact that today the needs of human beings are strongly influenced by technology, which has become a need in itself. We propose an analysis of how the Industrial Society has left the scene in favor of the Information Society, based on two meta-trends identified as Personalization and Distribution. Using an original biological metaphor, we describe “The DNA of ICT Evolution”, which uses the two meta-trends as filaments and ten trends as bases: ideal performance, ubiquity, flexibility, complexity, cognitivity, opportunist, cooperation, security, miniaturization, convergence. We then move on to a social analysis of the impact of the Information Society on nations and individuals, describing the Information Divide, composed by the Digital Divide and the Psychological Divide. We conclude the paper by arguing that a new skill will be needed by individuals to cope with all the above, which is the skill of disciplined creative thinking.

Index Terms—Information Society, personal needs, ubiquity, flexibility, complexity, cognitivity, opportunism, cooperation, security, miniaturization, convergence

I. INTRODUCTION: THE INFORMATION SOCIETY

Sociologists tend to agree on the fact that the main element that characterizes the present post-industrial society is the pervasiveness of Information, in its many possible forms and with all the associated operations of acquisition, storage, processing, exchange, generation. For the first time in history, life is forged by an intangible and conceptual element, and as a consequence it is clear that transformations and evolutions become much faster than in the past.

Understanding the world becomes therefore more difficult and more ephemeral, as transients and regimes are confused into a seamless evolutionary path. In addition, the comprehension of even the most basic operations on Information requires a considerable acquaintance with technologies, which we broadly define as Information and Communication Technologies (ICT), to include as a minimum the Internet, wireless and mobile communications, and the underlying micro-/nano-electronic devices. The necessity of a technical alphabetization has clear consequences on separating generations, and in generating the paradox that the younger and less experienced appears to understand more of the world, albeit only in terms of its technological texture.

In the Information Society, the tertiary sector of services gains a central position, taking a full stand alongside with the traditional, and more “solid”, primary and secondary sectors. Interestingly enough, ICT technologies have horizontal applications which also invest the primary and secondary sectors, such as for example the novel techniques for high precision agriculture, or the growing presence of ICT in manufacturing enterprises. In the Information Society, the tertiary sector has grown in centrality and criticality, to the point where a malfunction in the telecommunication infrastructure can paralyze an entire region.

To understand in depth this new form of society, it is required that we define the concept of Information more precisely. We elect to define Information as any conceptual entity that enables us to connect to the world and make sense of it. Three kinds of Information can be identified in general: historical information related to roots and origins, including the individual genome; social information derived from interactions with people and things, being part of a community; professional information consisting in the knowledge exploited daily in one’s working life. ICT technologies had their first impact on professional information, going back to the days when the Internet was game for a few, but then rapidly exploded into the realms of social and historical information, overcoming barriers and spreading a wave of homogeneity over the entire globe.

As a consequence of all the above, it can be concluded that access to Information is today a fundamental human necessity, which is required to satisfy basic personal needs. So the question arises as to what are the personal needs and how have they changed in accordance with this societal evolution. This paper attempts to address this issue in Section II, where we consider basic human necessities but with a technology oriented look, and we argue that technology itself becomes a need.

The attention is then moved onto technology, showing how it forges itself in relation with these new personal needs, giving life to new ICT trends. We argue that the two major forces that have driven the transformation from the Industrial to the Information Society are personalization, as opposed to standardization, and distribution, as opposed to the concentration of resources and intelligence typical of the Industrial Society. These are considered as meta-trends, which sustain ten specific trends, namely ideal performance, ubiquity, flexibility, complexity, cognitivity, opportunism, cooperation, security, miniaturization, and convergence. We dare proposing an original metaphor paralleling the biological DNA to the collection of the above meta-trends and trends in ICT. This can be found in Section III.

While the pervasiveness of technology in the Information society has definite and intrinsic positive aspects, it is also true that the impact on society, human relations, and even individual psychology is all but negligible. The fact is that there exists a new phenomenon which we identify as the Information Divide, which includes both the digital divide...
Fig. 1. A pyramid of personal needs with strong technological texture.

affecting all those who have no access to the Internet (or they access with insufficient bit-rate), as well as the psychological divide, related with the information overload which impacts upon all those who do have access. We dwell on these issues and their consequences in Section IV.

We conclude the paper by pointing to the one activity that we believe is instrumental in coping with the harsh consequences of the psychological divide: creative thinking, as the mechanism for becoming sources of information as opposed to mere sinks.

II. TECHNOLOGY AS A PERSONAL NEED

While it may not be desirable nor in the intention of the technologists themselves, it is true that humans get immediately used to new devices and services, to the point where they become real needs. This has been recently testified by the phenomenal take up of cellular communications, whereby anyone who forgets his mobile phone, or whose battery runs out, suffers definite psychological distress. The mobile phone has become a part of our bodies, and having it with us is definitely a need.

In view of the above, we want to review what can be considered to be the basic needs in today’s society, transformed under the light of technology. Starting from the individual is the preferred approach in order to avoid a brutal technology push. This is also undertaken in several policy programmes currently active in Europe, e.g. the Ambient Assisted Living (AAL) programme, which can be seen as the champion in focusing the attention on the human being [1].

Before we enter into the difficult realm of human necessities, we should underline the fact that we are not claiming the following discussion to be a psychological treatise on personal needs; rather, it is the interpretation of those needs in terms of their technological flavor. Re-interpreting in a more modern key the classical approach by Maslow [2], who modeled the personal necessities through a universally famous pyramid, user needs can be grouped into eight major categories as described in Fig.1.

Need 1 - Health

Health represents the most fundamental need of humans for truly satisfactory livelihood, without which any other asset loses its value. How can health be ensured through ICT technologies? This is the realm of e-health systems and technologies. The first application is to remotely assist patients, for example exploiting body area networks capable of monitoring vital parameters. E-health is the response to critical social challenges such as home care (or domiciliary care), i.e. healthcare or supportive care provided in the patient home by medical professionals, and independent living, i.e. providing assistance to disabled people. All of this is pursued at a significantly reduced social cost.

Once a complete e-health system will be in place, there is no doubt that it will be perceived as a crucial need, a necessity to have, and any malfunction will translate into serious problems. This will entail new levels of responsibility and accountability for ICT engineers.

Need 2 - Environmental fitness

Environmental fitness is a broad interpretation of the necessity for any human being to feel comfortable in the surrounding ambient. In our view, this includes two situations which have sufficiently distinct characteristics: fitting while in stationary conditions, and fitting in mobility, which opens the vast world of transportation. In terms of ICT technologies, these two dimensions can be respectively re-interpreted as the needs for ambient awareness and for smart transportation and info-mobility.

Ambient awareness integrates technology in two forms: hidden into the ambient itself (known as ambient intelligence, including all functions of domotics and sensor networks), or harnessed onto the person’s body to “feel” the context (known as context awareness). An important part of context awareness is the ability to know the position with sufficient accuracy, which indoors may require precision in the order of centimeters, as well as the capability to capture automatically information from the ambient itself.

Smart transportation systems include traffic monitoring and route optimization on one side (exploiting navigation systems), and info-mobility services to the travelers on the other, which can also be delivered in a peer-to-peer mode by exploiting vehicle-to-vehicle communications. Automatic driving is yet a dream, but it could arguably solve one of the most dramatic problems in today’s society, i.e. that of car accidents. Exploiting info-mobility systems, people can communicate, work, be entertained, and more generically reach the outside world while being on-the-move. Interestingly, ICT technologies can also reduce the need for mobility, by providing means for virtual interactions. This helps in exploiting resources efficiently and in favoring sustainability, in line with the Green ICT smart energy requirements [3], [4], [5].

All of the above depicts a situation whereby all the required technologies are absolutely necessary to feel fit in one’s environment.

Need 3 - Security and Privacy

Security and privacy are primary needs of any person to feel protected during any aspect of life.

This necessity translates into the need for preserved secrecy on private data, which becomes a real social challenge in an era where communication of personal digital data is at the center...
of many ICT applications, especially with the growing diffusion of Web-based Internet applications and social networks. Straightforwardly, in the ICT context this need translates into putting in place strategies for cybercrime prevention, avoidance of personal identity theft and dissemination of malware and spam; all these objectives can be reached via enhanced authentication mechanisms and digital identity management.

With the success of Internet and the ever increasing individual exposure over the net, the need for secure, certificated, authenticated transactions will become unavoidable.

Need 4 - Social Interaction

Social interaction is the natural need of humans to belong to some form of society, from a small scale (a family, a group) to a larger scale (a community, a Country), in response to the continuous search for support, discussion, comparison and opinion exchange with other persons.

This need takes a completely new connotation with the recent technology advances. In particular, it is worth mentioning here the tendency for social networking, so highly popular nowadays amongst younger generations, and the trend towards virtual interactions in cyber-realities where all aspects of life can be artificially re-constructed, through the creation of virtual identities or avatars (e.g. Second Life). The web is now used in a participating mode, as testified by the success of a number of applications such as Wikipedia, blogs, MySpace, Facebook, Twitter, YouTube, GoogleMaps, etcetera. In this framework, other foreseen technologies are enhanced reality (with brain electrochemical stimulation), and augmented reality (combination of real and digitally modified identities and environment).

Even if we believe that real interactions among human beings should not ever be replaced by their artificial counterparts, social networking, enhanced reality, and augmented reality are gaining widespread diffusion, which is likely to grow significantly in the medium-term.

Need 5 - Information and Learning

Information and learning is the natural need for knowing, being up-to-date and informed anytime and anywhere.

This need has assumed a particularly compelling connotation in recent years with the rapid and capillary diffusion of the Internet, via the use of search engines and the introduction of the semantic web. In the digital era, the person is indeed more and more eager to reach rapidly and efficiently all contents in a precise moment, at home, at work, on the move. Besides the Internet, e-learning and digital libraries are other examples of ICT technological responses to this need, both contributing to satisfy the user thirst for receiving information and learning in remote areas.

Notably, these trends are first responses to the social challenge of inclusion, which contributes in avoiding person isolation from society through e-inclusion mechanisms. As broadband becomes a necessity of daily life (according to the “broadband for all” paradigm), the impact of information exclusion for citizens that do not have broadband access or who cannot afford it will be dramatic. Today’s digital divide may become tomorrow’s social-exclusion. This fundamental social problem will be discussed again in Section IV.

Need 6 - Working Life

Working life is the need for self-fulfillment and self-achievement realized through working activities, which enables to exploit talents and education, contributing fundamentally to self-esteem.

From the ICT technological point of view, this need is reflected in the recent e-business applications, which comprise for example home-based or mobile-based tele-working, broadband connection between different enterprise premises, remote training, supply chain management, electronic orders processing, customer service via Internet.

The nomadic use of ICT technologies will certainly challenge the meaning of ‘being at work’, and e-business will become the basis of the future working life paradigm.

Need 7 - Transactions

Transactions is the need of humans to be supplied with services and goods in order to satisfy material/immaterial desires.

This general need is reflected into ICT technologies through e-commerce, which consists of buying and selling products or services over electronic systems such as the Internet. E-commerce is growing considerably in business, private, and consumer contexts. Home banking, on-line shopping, electronic tickets are only some examples.

It is already a fact that e-commerce is a daily life tool for transactions, and it is almost a certainty that it will become more and more diffuse in the medium-term.

Also, e-government applications and services are gaining importance in everyday life of citizens to simplify bureaucracy, drastically reduce public administration costs and bridging the age divide by bringing public administration services directly into citizens’ homes.

Need 8 - Entertainment

Entertainment is the need for amusement and hobbies, to aliment the innate human tendencies towards recreation.

ICT is strongly reflected in this need. In partial overlap with the social interaction need when applied to the personal sphere, the necessity for entertainment is one of the most typical goals of ICT, taking on various forms such as on-line gaming, portable consoles, three-dimensional television and cinema, mobile television, peer-to-peer downloads, MP3 players, and social networking.

We can state that the diffusion of these technologies will become more and more essential for individuals, providing new ways of experiencing entertainment, with quantum leaps with respect to traditional ways of being entertained.

III. MAJOR TRENDS IN THE INFORMATION SOCIETY

Taking on an original point of view, here the underlying and unifying major trends in the Information Society arena are presented. As already discussed in the previous Section,
ICT applications provide several ways to satisfy personal needs. However, we want to go beyond a mere listing of ICT keywords, but rather point at extracting the major trends which lie within all the above mentioned technologies and their future evolutions.

We elect to say that the two main forces which are governing the Information Society revolution are personalization and distribution. Using a biological metaphor, we can think of personalization and distribution as the two strands, the two filaments, in a DNA structure, upon which and through which other trends are formed. This is what we call the DNA of Information Society evolution, represented pictorially in Fig. 2. The two filaments generate and are linked by the bases, each one corresponding to a major Information Society trend. Therefore, in the following personalization and distribution will be identified as meta-trends, or “trends of trends”.

As in the DNA structure, the two fundamental strands of personalization and distribution are actually running into opposite directions, i.e. they are the anti-parallel support of the Information Society DNA double helix. In fact, while personalization implies a local view, distribution naturally translates into a global reality. Therefore, Information Society is imposing on the world a complex organization, whereby the global truth is formed through belief propagation from individual nodes which intelligently work within their local boundaries.

A. The personalization meta-trend

The success of the industrial society was based on the mass production of goods at low cost, to be sold to a consumer market with homogenized tastes and desires. This paradigm today is completely reversed. The strategy is to go to the person, produce for the individual, satisfy specific needs, segment the market into small niches each tailored to a particular group of persons. This is all made possible by the fact that today it is feasible to produce at low cost with flexibility and modularity. This trend is clearly reflected in the world of communications. Since the advent of mobile telephony, also identified as personal communications, users do not call a location, but an individual. Users are now used to see the number that is calling, associated to a person in their phonebook, and can freely decide whether to take the call or not. Users can select a specific tariff structure, suited to their calling profile, and they may even have lower tariffs when they call specific groups of people. In terms of accessing the Internet, users can start from their preferred home page, can browse over portals shaped according to their own profile, can select their favored content from a huge offering, can even become content producers by uploading pictures and movie clips. The individual has become a source of information, and not only a sink to be filled with advertisement. Even when watching TV, users now have at their disposal a growing number of on-demand offerings, from which they can freely select. The individual is more important, enjoys more freedom, is much more active than it was in the past.

B. The distribution meta-trend

Another fundamental ingredient of the industrial society was the concentration of resources and intelligence into a few centers. This made interactions and investments more efficient, while also increasing the risk associated to losing a center, and caused large movements of people (the first of which was the move out of the country into cities). The structure of companies was based on rather rigid pyramids, with very specific work functions and descriptions, and linear work flow procedures. Again, also this paradigm is reversed in the information society. Today, there is a clear trend towards the distribution of resources. In the world of manufacturing, it is customary to have parts produced in geographically distant premises. This reduces costs and creates a global economy where effects propagate unboundedly. Intelligence is distributed and decisions are obtained through a network of interactions. This adds greatly to the responsibility of each person in the organization, and the work functions become ever more flexible. This also creates a need for continual education, for it is not possible to adapt to the fast dynamics of the current societal evolution if one considers that his/her training ended in school. Companies’ structures evolve from pyramids into networks of intelligent nodes, and the structure may evolve on a per product basis. The more futuristic version is the virtual company, where different entities join into specific ventures which only have the lifetime of a product life cycle.

It can be stated, without the minimal shade of doubt, that all of this is resting upon Information Society technologies. The network of people relies on the telecommunication network, and, in the future, the same will happen for the network of things, also known as machine-to-machine (m2m) communications [6]. A distributed society and a distributed company need to find the necessary information anywhere they might reside. This is only possible thanks to the Internet and the associated search engines which allow to have nearly all the information in the world at one’s fingertips, also in smart ways through exploitation of the Semantic Web (e.g. Bing) [7]. Intelligence and storage are pushed to the edges (think for example about cloud computing), which reduces the risk associated to the loss of any network node, but at the same time requires a new ethical code, as well as security in communications. Networks evolve from heavy infrastructures to lightweight ad-hoc self-organizing topologies, where the role of operators needs to be defined anew. Software itself becomes a distributed resource, now intended as a service and not as a product.
C. The ICT trends

The two meta-trends of personalization and distribution can be combined in various ways to form the trends of ICT technologies, which according to the DNA metaphor correspond to the bases, the sequence of which identifies the genetic code.

The following is our elected list of ten trends: Ideal performance, Ubiquity, Flexibility, Complexity, Cognitivity, Cooperation, Security, Miniaturization, Convergence.

1) The ideal performance trend: The search for the ultimate ideal performance is the major force behind the evolution of any technical or technological system. It appears to be in the nature of the human kind to strive for the extraction of the maximum possible output, the optimal exploitation of resources, with the largest possible efficiency. This requires knowledge of the ultimate performance boundaries. In the case of communications, the boundaries are the object of Information and Communication theories, which in many instances do indicate where these limits are. The trend is therefore towards achieving the performance limits set by Information Theory. How does this link to the two meta-trends? The optimization of performance requires perfect fit to the specific communication conditions (propagation channel, interference, transmission format, etcetera), which can be interpreted as matching the individual user needs and constraints. This is part of personalization. On the other hand, the limits set by Information Theory are not restricted to a single link, but can and should be extended to the more complex case of networks. In this case, achieving the optimal limits requires global optimization, to balance fairness and overall throughput, with distributed intelligence. This is evidently part of the distribution meta-trend.

2) The ubiquity trend: The trend towards ubiquitous communications, the overused “Anywhere, anytime” motto, has been the driver for the evolution of cellular communications since the 70’s. Coverage is today extremely good in most urban areas, and surprisingly good in unexpected locations, even though obviously gaps remain in developing parts of the world. What is yet necessary is to sharply increase the geographic spectral efficiency (in bit/s/Hz/km²), to provide ubiquitous broadband wireless access. Associated to ubiquity, we find mobility and pervasiveness. Mobility is the trend towards communication systems which can interconnect terminals moving at any speed, including all types of vehicles, trains, airplanes, ships. Pervasiveness is the trends towards finding connectivity all around us, in a truly wireless ambient. Even power supplies can become wireless, with free charging points distributed is strategic locations. This all adds to the trend towards “living without cables”.

Smart cities [8] are a consequence of this trend, addressing ubiquitous connectivity within a city intended as an IT-district, towards the concept of the Internet of Citizens. This concept has many important social implications. Essentially, by living into a collaborative wireless ambient, the individual can benefit from optimized environmental fitness, which satisfies a basic human need. This is clearly a specification of the personalization meta-trend. At the same time, this personal fitness can be carried along in any location, becoming ubiquitous fitness, an evident derivation from the distribution meta-trend. Even though ubiquity and distribution may seem very similar concepts, we separate them by limiting the interpretation of ubiquity to the pervasiveness of wireless networks, and by attributing to distribution this and all other implications related to social aspects, work organization, globe economy, etcetera.

3) The flexibility trend: Along with the search for Ideal Performance, this is also a major trend in the evolution of any technical system. All engineering systems start as rather simple and rigid, performing but a few functions, with limited scope for modifications in response to user needs. In the course of its development, the system acquires more and more functions, more and more options, which can be selected flexibly depending on instantaneous necessities. This is a very strong trend in wireless communications. Transmission systems, protocols, and terminals are being designed as reconfigurable entities, with capabilities that can be flexibly adapted to the conditions set by the propagation channels, the transmission buffers, the spectrum availability, the interference environment, the desired quality of service, etcetera. Dynamic spectrum assignment strategies are being devised and starting to find their way into regulatory policies. Digital electronics capabilities are exploited to design software radios and flexible radios. Even analog electronics is now being bent to the requirements of designing flexible Radio-Frequency (RF) front-ends, with reconfigurable filters over large bandwidths. It is an easy task to map the trend towards flexibility as a direct son of the personalization meta-trend. In fact, it is obvious that flexibility is only useful if it is used to accommodate individual conditions and needs. On the other hand, it may be harder to describe the connection with the distribution meta-trend. However, flexibility at system level requires knowledge of all local conditions, in order to find a global optimum satisfying the requirements of the entire user population. Therefore, we can say that global flexibility is related to the trends towards distribution of intelligence, where as a minimum each user must sense its own environment and feed back this information to peers or to base stations. Also, the trend towards flexible network topologies is clearly out-spinning from the distribution meta-trend.

4) The complexity trend: Technical systems always evolve towards increasing levels of complexity, as functionalities increase and performance improves. On the other hand, technological complexity can become a major hurdle in its usability. Therefore, while internal complexity increases monotonously, there is a contextual trend towards the simplification of the human-to-machine interface. Complexity and simplicity must live together in harmony. As complexity grows, we must face the danger of increasing energy consumption, which could make entire systems unsustainable. A clear trend towards the design of “green” technical systems is growing powerfully nowadays [3]: we could say that energy consumption is to complexity as energy saving is to simplicity. The long-term sustainability of human activities will be one of the paramount technological challenges of the 21st century. Since the data traffic in communication networks is exponentially increasing, energy-efficient communication techniques are needed for as-
suring that communication-related energy consumption is not exploding and that the pertinent carbon footprint is caped or even reduced [6]. In order to meet global and national goals for carbon-footprint reduction and to compensate for noticeably increasing energy expenditures, technological energy-saving measures are a mandatory ingredient of any emergent Information Society technology, be it in the short- or long-term, be it for evolutionary or revolutionary technology.

Mapped onto the world of wireless communications, complexity is visible in systems, protocols, terminals, network equipment, essentially in every element. The need to simplify is stringent for user terminals, but also for network management. And we can say that “green” communications are emerging as a very hot area of research and development. The relationship between complexity and the personalization meta-trend is inherent in the fact that we do not accept standard and rigid solutions, but rather we always look for configurations which are adapted to individual needs. A personalized solution is always more complex than a standard item. The key enabler for the realization of complex systems is the fact that today we are able to produce personalized objects in a very cost effective manner. It is also true that, in many instances, personalization is perceived by the final user, but it is in reality a specific combination of a few standard objects. In view of the distribution meta-trend, it should be apparent that distributing intelligence, responsibilities, management functions, all translate into a more complex system. In this case, complexity also brings in the concept of emergence: the arising of novel and coherent structures, patterns and properties during the process of self-organization in complex distributed systems. Emergence can be weak when it can be reduced to its elemental parts, or strong when irreducible. Irreducible emergence can be thought of as an independent system, living a life of its own.

5) The cognition trend (including also self-organization and bio-inspiration): As complexity of systems grows larger, control becomes more and more difficult. At a first inspection, it would be desirable to be able to set rigid rules to which all system elements should abide. This has worked in the past and still does today. However, this can only be pushed to a limit, when exceptions to the rules become frequently passing, to the global behavior. This can be brought to the extreme where the overall objective functions, such as for example the estimation of a parameter, are elaborated only in a distributed manner, and the final result is not necessarily collected at a fusion center, but can itself be distributed into the network.

6) The opportunism trend: With increasing degrees of distributed intelligence, flexibility, and complexity, it becomes crucial to execute operations not at any generic time instant, but when and only when the conditions are optimal. In other words, it becomes necessary to catch the opportunity for performing a specified task in the most efficient manner, and with the largest associated benefit. Indeed, all systems are dynamically varying in several dimensions (as a minimum, time), which means that conditions will fluctuate and opportunities will be created. To use a dimension or another, depending on the underlying conditions, can be interpreted as a form of diversity. Therefore, opportunism is a way to exploit diversity, choosing from time to time the path which offers the minimum resistance to our action, and thus optimizing the use of resources. In a sense, opportunism can be seen as the opposite of the brute-force approach, where exploitation of resources is total and completely independent of the ensuing conditions. The beauty of this is the fact that, in a network (be it technical or social) the total amount of resources is limited, so if each one use the minimum necessary to achieve its own purposes, then the overall efficiency is maximized. In other words, the use of brute force from any single individual hurts the entire network. In wireless systems and networks, and particularly in the family of Beyond 3G cellular networks, opportunism has become a major flagship for resource assignment, scheduling, multiple access. Resources are given dynamically to those terminals which are at a particular instant enjoying the best channel conditions, which will allow to serve them with the minimum effort and maximal efficiency. In order to avoid that some terminals are always left out of the game, opportunism should always go along with fairness, implemented in one of its several possible embodiments. In this specific case, the personalization meta-trend materializes in the fact that we go after the opportunity which is occurring for a specific individual, knowing that it will only last for a limited window of time. On the other hand, we want to be fair to all users, and as such protocols and strategies are ready to consider also the needs of those for which opportunities do not seem to happen, at least not with sufficient frequency. In terms of the distribution meta-trend, we observe that opportunities may also be visible at a local level. This is because, to enable scalability, it is not conceivable that all information be collected in a
single decision making node. Therefore, decisions to seize specific opportunities should be taken locally, with feedback on instantaneous conditions transmitted only when and where necessary, possibly on short legs to minimize latency and thus maximize network reactivity. Hence, opportunism must go along with distribution.

7) The cooperation trend: Cooperation can be seen as the virtuous consequence of awareness. If an individual, or an entity, is isolated, it can only work for its own specific goals. On the other hand, even if the entity is not isolated, but is unaware of the needs, or even the sole presence, of other entities around it, it will behave exactly as it did in isolation, working undividedly towards the achievement of its objectives. Only when an entity becomes aware of the presence, requirements, and needs of other entities around it, then it can realize that working in isolation may not be the most efficient way. Even the objectives are modified, at least because one sees not only its own objectives but also those of others, thus creating the notion of global objectives. Awareness generates a change of perspective, which can lead to various forms of cooperation amongst the individuals. Cooperation requires trust, fairness, and regulation, in order to ensure that all individuals benefit from the process. Cooperation in wireless communications can, for example, take the form of relaying the information sent by another user, in order to help it reach the final destination. In this way, the cooperating node is spending part of its resources not to achieve its own objectives, but rather to help another node do so. In return, it will trust that the situation will reverse when its own opportunity comes along. It is clear then that cooperation and opportunism go together, as the mechanism for cooperation will adapt itself to the underlying conditions which will vary dynamically over time. Other interesting forms of cooperation can be envisaged for virtual beamforming, virtual MIMO, collaborative positioning, etcetera. Cooperation is an act between individuals, and as such it possesses intrinsically the character of the personalization meta-trend. The personalized network entity is aware of the other entities, cognitively decides that it is useful to cooperate, trusts the other entities, expects to receive mutual benefits and to achieve its own goals while contributing to the global goals. This comes very close to the description of the behavior of a person in a social network. On the other hand, the exploitation of cooperation means, once more, that the operations in the network do not belong to a single terminal and a single central control entity, but rather require the involvement of a multitude of actors, distributed of the area of service, whereby decisions and operation occur as the result of the overall interaction. This is clearly in line with the distribution meta-trend, and we can say that cooperation without distribution is impossible, and distribution without cooperation is less effective and not exploited to its fullest.

8) The security trend: We must also recognize that, in front of all the positive aspects brought in by the personalization and distribution meta-trends, there is also an associated increase in the risk of misuse of Information Society technologies. Centralized control may be bulky and in some cases unfair, but it can also serve as a guarantee for secure transactions, which can be protected more easily by various kinds of threats. On the other hand, when organizations become distributed, when decision making is the result of consensus, when resource management requires information from the edges, then it is clear that there are so many more possibilities for an alien to come in and disturb or deviate the process far from its intended objectives. And since there is a trend for personalization, any individual or any entity is up front with all of its features, which can be stolen or misused in many ways. Therefore, the meta-trends of personalization and distribution require that much attention is paid to ensure security, guarantee privacy, defy malicious attacks, propagate trust. This applies to society in general, and certainly it does also to wireless communications, which traditionally have been the weak side of network security. One special word for trust: it is not just a matter of making sure that content is encrypted, that access is conditional to authentication, that sensitive data is not exchanged (or at least not frequently). It is also a matter to make sure that the final user perceives that using Information Society technologies is secure. In other words, there must be trust in Information Society technologies, or else the uptake will always be below expectations, and the impact much more limited than the potential.

9) The miniaturization trend: In the evolution of technology, we always see a trend towards miniaturization, as the results of improvements in the processes and in the understanding of the underlying physics. This has held marvelously in the case of digital electronics, where the scale of integration of ICs (Integrated Circuits) has grown exponentially through the years. Digital ICs are horizontal enablers for the progress in wireless communications, not only from the technical point of view, but also from the economic side, given that the cost of ICs has also decreased steadily through the years. And the end is not in sight: with the rush for nanoscale devices, unprecedented improvements are yet on their way. Also, circuits built on organic materials promise to change forever the notion of "hardware", as we will be seeing and use devices in plastic and even softer materials. Nanotechnologies will produce entire nanosystems, which can be distributed as smart dust for a multitude of distributed sensing applications. Here comes the relationship with the distribution meta-trend. On the other hand, the trend towards miniaturization can also be seen in a more general way. In terms of cellular mobile networks, there is a clear trend towards the miniaturization of cells, which went from macro to micro, pico, and now femto-cells. The femto-cell is intended to be installed by and individual in a home or a small office, and as such we can see the connection with the personalization meta-trend. We can also see the miniaturization of networks, as for the case of the body area network, interconnecting different parts of the body for various applications for the benefit of the individual, such as e-health.

10) The convergence trend: Last but not least, convergence. It is placed at the end because in a way it connects all previous trends. In a general sense, convergence can be seen as that process according to which concepts which used to be separated come together to form new meaning. The process eliminates barriers and distinctions, and creates larger classes. Since we use classification as an instrument for clarification, it
is indeed true that convergence always causes a certain amount of confusion, as previous certainties are questioned and new approaches must come in. In economic terms, convergence is an earthquake that shakes market shares and inevitably increases both opportunities and threats. From the point of view of scientific research, convergence can be seen as the uprising of interdisciplinary research, where competences from different areas are merged. The major benefit is that different frames of mind coming together have the potential to produce breakthrough innovation. From the point of view of wireless communications, convergence plays a major role in at least two ways. First of all, the distinction between mobile and fixed telephony is vanishing, with operators offering bundles which include also Internet access and TV (the so-called quadruple play). Considering also the fact that IP (Internet Protocol) is rising as the common network protocol for all services, we can see clearly the incredible force of service convergence. Secondly, there is also convergence in the world of wireless terminals, which more and more become phones, computers, cameras, organizers, etcetera. Let’s interpret the trend for convergence in view of the personalization meta-trend. No matter where the person is, we want to provide the same access conditions, the same service profile, as if the person was virtually at home. Convergence of networks and terminals can enable this concept. On the other hand, this can also be seen as the famous “anywhere, anytime” motto, i.e. the fact that we are always surrounded by a converged (albeit heterogeneous) network infrastructure, of which we don’t want to know the details, as long as we can use it for our purposes. Therefore, convergence can be seen as an enabler for the distribution meta-trend, because without it we would not see a seamless wireless ambient but rather a jigsaw puzzle of technologies.

IV. SOCIAL IMPACT: THE INFORMATION DIVIDE

The pervasiveness of technology in the Information Society presents undubitable benefits in terms of quality of life. However, there are also costs and negative consequences, which cannot and should not be neglected. The engineering point of view must here subside in favor of sociological considerations.

Let us imagine for an instant that the provocative assumption “all of the world’s knowledge will be on the Internet by 2020” holds true. Although no one will ever be able to prove this, it is a fact that we are witnessing incredible growth in both the number of new web sites opened daily [9] and in Internet traffic (forecasts by Cisco estimate more than half zetta-byte of global traffic to be reached by 2012 [10]). It is also becoming a habit to verify facts and notions by matching what we know with the results of a web search. It is clear that, under the above assumption, the possibility to access the Internet becomes a major discriminator for individuals, groups, and even entire nations, creating a fundamental issue which we here identify as the Information Divide, intended as the separation between one entity’s knowledge and the rest of all information. The information divide is not limited to a lack of technology infrastructures, because it can exist even with full access capabilities. Indeed, the information divide can take on two totally different, or dual, forms: the digital divide and the psychological divide.

The digital divide, i.e. the lack of proper telecommunication infrastructures, has critical impact both in professional and social terms, which can be unbearable for the younger generations. We can distinguish two kinds of digital divide: geographical and demographic. The geographical divide applies between continents and between specific countries within a continent. The geographical divide especially affects developing countries, where the benefits of the Information Society revolution are less clear, to the point where the gap with the developed countries tends to be exacerbated. Focused efforts and investments are needed to transform the divide into a digital opportunity [11]. The demographic divide refers to the clear difference in quality of service and coverage between urban areas and areas which are less densely populated, broadly defined as rural areas. The problem is one of economy of investment, and it cannot be expected that private entities take on all of the costs associated to giving full connectivity to rural areas. Again, institutional effort is required: bridging the digital divide requires large investments in telecommunications infrastructures, which can range from the cost-effective ADSL solutions to the super-high capacity optical fibers, from broadband cellular systems (which enable high penetration) to broadband satellite links (providing large coverage in remote areas).

The second type of divide is the much less explored psychological divide, which is related to the inability to handle information, due to either information overload or illiteracy in new technology matters. The psychological divide is related to age and education, as well as to the inability to handle huge amounts of information. The generational divide relates to the fact that younger generations are more used to technology with respect to older people. In fact, they are born with technology, and have a hard time imagining a world without it. It is a fact that having children in a household largely increases the demand for having Internet access. We are facing the paradox whereby young inexperienced people understand more of the technological world than older and wiser people, with an impact on familiar and social relationships. The educational divide derives by differences in the scholastic background. Several studies have confirmed that a small percentage of the people with lower secondary education access the Internet, while the percentage rapidly grows to saturation for people with higher and higher education. Finally, there exists a psychological problem in dealing with instantly available infinite information. While it is possible to download virtually unlimited amounts of data on any subject, it is undubitable that knowledge without experience is not wisdom. The progress towards the semantic web is a step towards pre-filtering of information in terms of its relevance, which should help considerably in reducing the information overload. But will this be sufficient? And from the professional point of view, what is the impact of the fact that the know-how will not be a major discriminator in the future?
V. CONCLUSIONS: A CREATIVE FUTURE

The Information Society is such that technology is pervasive and a true necessity for individuals, which depending on their condition are faced more or less fiercely by the Information Divide. We believe that the impact of the psychological divide may be dramatic on some individuals, who may be overwhelmed by the feeling that learning anything has much reduced relevance when all information is available at your fingertips, and they may reduce to living their life in a very passive mode. Entire lives could be spent simply enjoying online content.

Fortunately, there is a way out, which however requires a true paradigm shift in terms of the way we approach thinking and education. Throughout our history, we have always held at a prime the skill of critical thinking, i.e. the capacity to judge and criticize facts, ideas and happenings with respect to our knowledge. This is and will always remain a fundamental asset for any individual, in order to live a proper life. However, this will not be sufficient in the future evolution of the Information Society, because critical thinking does not generate new knowledge, rather it is a powerful filter that preserves our culture and our roots. We need to add the skill of creative thinking, intended as the capacity to generate new concepts and ideas in a disciplined and controlled manner.

Creativity is not yet a science, although there are many sciences that deal with creative thinking, such as the history of science, the philosophy of science, psychology, engineering, to name a few. But we are still missing a coherent view and a scholarly approach, which is necessary to turn creative thinking into an educational subject, taught in schools and universities as a fundamental and transversal skill. But we strongly believe that creativity will become a general subject and will grow to become the principal tool to allow humans to deal with the psychological divide imposed by the Information Society. Only when one becomes a source of new ideas then it is possible to come to terms with the information ocean.

ACKNOWLEDGMENT

This work has been supported by the EC FP7 NEWCOM++ project.

REFERENCES


Giovanni Corazza is a Full Professor at the University of Bologna, Head of the Department of Electronics, Computer Science and Systems (DEIS), and leader for Wireless Communications inside the Advanced Research Centre for Electronic Systems (ARCES). He was Chairman of the School for Telecommunications in the years 2000-2003, Chairman of the Advanced Satellite Mobile Systems Task Force (ASMSTF), Founder and Chairman of the Integral Satcom Initiative (ISI), a European Technology Platform devote to Satellite Communications. Since 1997, he is Editor for Communication Theory and Spread Spectrum for the IEEE Transactions on Communications. He is author of more than 200 papers, and received the Marconi International Fellowship Young Scientist Award in 1995, the IEEE 2009 Satellite Communications Distinguished Service Award, the 2002 IEEE VTS Best System Paper Award, the Best Paper Award at IEEE ISSSTA’98, at IEEE ITC2001, and at ISWCS 2005. He has been the General Chairman of the IEEE ISSSTA 2008, ASMS 2004, ASMS 2006, ASMS 2008, ASMS 2010 Conferences. His research interests are in wireless and satellite communications, information society, estimation and synchronization, spread spectrum and multi-carrier transmission, upper layer coding, navigation and positioning.

Alessandro Vanelli-Coralli received the Dr. Ing. Degree (cum laude) in Electronics Engineering and the Ph.D. in Electronics and Computer Science from the University of Bologna (Italy) in 1991 and 1996, respectively. In 1996, he joined the Department of Electronics, Computer Science and Systems (D.E.I.S.) at the University of Bologna where he is currently an Assistant Professor. Since 2001, he has been a Research Associate of the Advanced Research Center for Electronic Systems (ARCES) of the University of Bologna. During 2003 and 2005, he was a Visiting Scientist at Qualcomm Inc. (San Diego, CA). He participates in national and international research projects on satellite mobile communication systems and he is Vice-Chairman of the R&D group of the Integral SatCom Initiative (ISI) technology platform. His research interests are in the area of wireless communication systems, physical layer techniques, and digital signal processing. He is a member of the Editorial Board of the Wiley International Journal of Satellite Communications and Networking and he has been co-editor of several Special Issues on Satellite Communications in International Scientific Journals. Dr. Vanelli-Coralli is the General co-chair of the 4th ASMS2010 and he has been the Technical Program Committee Chairman for the 4th IEEE IEEE ASMS2008 and Vice-Chairman for the IEEE ISSSTA 2008 Conference. Dr. Vanelli-Coralli co-authored more than 130 papers and scientific conference contributions and is co-recipient of the Best Paper Award at IEEE ITC 2001 and at IEEE ISWCS 2005. He is an IEEE Senior Member.

Raffaella Pedone was born in Bologna (Italy) in 1976. She received the Dr. Ing. degree (summa cum laude) in Electronic Engineering in 2002, and the Ph.D. in Telecommunications, Electronics and Computer Science in 2006 from the University of Bologna (Italy). She is currently a Post-Doc Researcher at the Department of Electronics, Computer Science, and Systems (DEIS) and the Advanced Research Centre for Electronic Systems (ARCES) of the University of Bologna. Her research interests are in the field of communication and information theory, with particular emphasis on synchronization techniques for spread spectrum systems and TDMA networks, and interference mitigation techniques. She is involved in several national and international research projects on satellite mobile communication systems, and satellite/terrestrial convergence. She is author of more than 30 papers published in International Journals and Conference Proceedings. She was co-recipient of the Best Student Paper Award at the international conference IEEE ISSSTA 2004, Sydney, Australia, and of the Best Paper Award at the international conference IEEE ISWCS 2005, Siena, Italy.
ADVANCES in Electronics and Telecommunications is a new peer-reviewed journal published at Pozna University of Technology, Faculty of Electronics and Telecommunications. It publishes scientific papers addressing crucial issues in the area of contemporary electronics and telecommunications. Detailed information about the journal can be found at: www.advances.et.put.poznan.pl

During the last years telecommunications techniques made a revolutionary step forward. Such a progress was stimulated by growing requirements from traffic being served in modern networks. Broadband service platforms offer a significant leap in types of information available to end users, offering a multitude of new services. Video on demand, multimedia, unlimited Internet access become also possible in wireless systems. There are many questions that need to be answered including: how to interwork existing services with future wired and wireless networks, how to manage and optimise the transmission of growing traffic volumes, how to enhance networks performance. To discuss aforementioned topics (and probably to give rise to many others) the following areas have been selected as the primary categories for article submission:

- Teletraffic Models and Traffic Theory
- Performance Evaluation and Modeling of Communication Protocols
- Traffic Issues for Mobile Systems
- Traffic Issues for Internet
- Traffic Issues for High Speed Networks
- Traffic Models for Optical Networks
- Routing and Wavelength Assignment in Optical Networks
- Switching Fabrics and Node Architectures, Optimization, Evaluation and Modeling
- Performance Analysis Methods and Simulation
- Traffic Measurements
- Intelligent Routing Protocols
- Broadcast and Multicast Traffic Control and Management
- Traffic Control and QoS
- Network and System Architectures
- Mobile Networks
- Network Design and Planning
- Traffic and Network Management
- Computer Systems and Networks
- Quantitative Measures

The special issue of the Advances in Electronics and Telecommunications will collect original and unpublished research contributions in the area of teletraffic. The authors should submit their contributions electronically to: www.advances.et.put.poznan.pl/?p=authors. The authors are asked to submit their papers in English in accordance with the IEEE Transactions style defined at the website: www.advances.et.put.poznan.pl. The editor reserves the right to edit the contents of the paper if language and style still needs some improvements.

Important deadlines:
- Manuscript submission: September 1, 2010
- Acceptance notification: November 1, 2010
- Final manuscript due: December 1, 2010
- Publication: February 1, 2011
ADVANCES in Electronics and Telecommunications is a new peer-reviewed journal published at Pozna University of Technology, Faculty of Electronics and Telecommunications. It publishes scientific papers addressing crucial issues in the area of contemporary electronics and telecommunications. Detailed information about the journal can be found at: www.advances.et.put.poznan.pl

The aim of this special issue is to provide an overview of the recent advances and the state-of-the-art of image and video telecommunications. The topics of interest include, but are not limited to:

- new image and video compression techniques,
- image and video encoder/decoder implementations,
- video streaming, scalable video transmission,
- networked video services,
- 3D video,
- image and video watermarking,
- error concealment in image and video,
- image and video enhancement and restoration,
- UHDTV technology,
- international standards in image and video communications,
- image printing and video displaying,
- image and video quality assessment,
- image and video content retrieval.

Prospective authors are invited to submit their contributions electronically to: advances@et.put.poznan.pl. Authors should indicate that their submission is intended for this special issue. The authors are asked to submit their papers in English in accordance with the IEEE Transactions style defined at the website: IEEE Transactions style. Please note that submission will be treated as a declaration that the manuscript contains original material.

Important deadlines:
- Manuscript submission: December 20, 2010
- Acceptance notification: April 10, 2010
- Final manuscript due: May 1, 2010
- Publication: July, 2011

Final call for papers will be published on advances@et.put.poznan.pl in September 2010.
About the journal

ADVANCES IN ELECTRONICS AND TELECOMMUNICATIONS is a peer-reviewed journal published by Poznań University of Technology. It publishes scientific papers devoted to several problems in the area of contemporary electronics and telecommunications. Its scope is focused on, but not limited to the following issues:

- electronic circuits and systems,
- microwave devices and systems,
- DSP structures and algorithms for wireless and wireline communication systems,
- digital modulations,
- data transmission techniques,
- multiple access techniques and MAC issues,
- information and channel coding theory and its applications,
- software defined radio and cognitive radio technologies,
- wireless local area networks (WLANs),
- satellite communication,
- navigation and localization,
- synchronization subsystems,
- time and timing,
- modeling techniques of package & on-chip interconnects,
- radiation & interference, electromagnetic compatibility,
- propagation aspects in wireless communication,
- UWB channel modeling,
- measurements and wireless sensor networks,
- web technologies,
- e-learning,
- multimedia communication,
- audio and speech processing,
- image and video processing,
- software and hardware system implementation,
- advanced A/D and D/A conversion techniques and their applications,
- SDI - Software Defined Instruments,
- effective measurement, estimation and computation of signal parameters,
- consumer electronics.

Detailed information about the journal can be found at: www.advances.et.put.poznan.pl

The Editorial Board invites paper submissions on the above topics for Open Call.