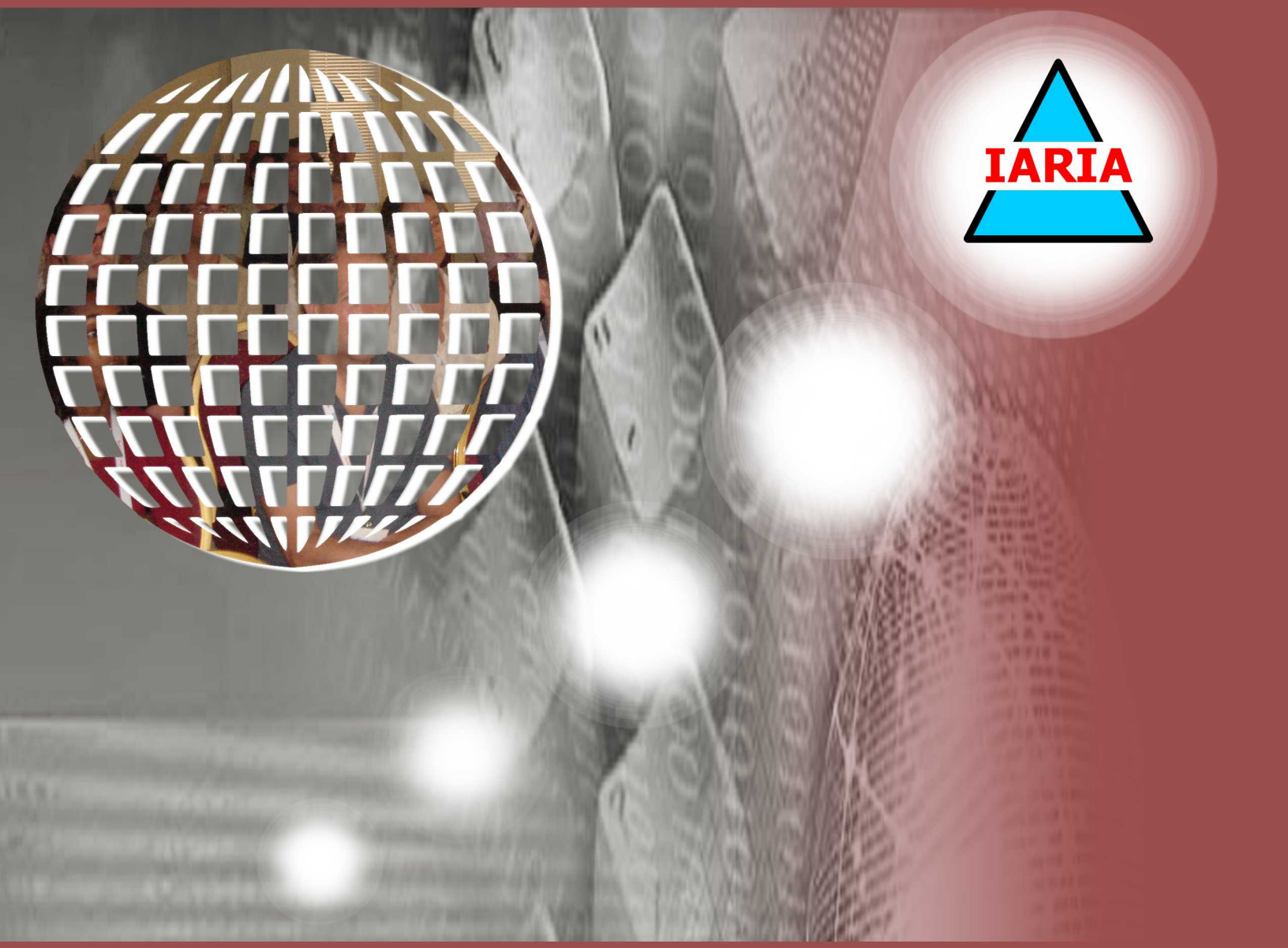
International Journal on Advances in Telecommunications



2015 vol. 8 nr. 1&2

The *International Journal on Advances in Telecommunications* is published by IARIA.
ISSN: 1942-2601
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Telecommunications
Volume 8, Number 1 & 2, 2015

Eugen Borgoci, University "Politehnica"of Bucharest (UPB), Romania
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Christos Bouras, University of Patras, Greece
Martin Brandl, Danube University Krems, Austria
Julien Broisin, IRIT, France
Dumitru Burdescu, University of Craiova, Romania
Andi Buzo, University "Politehnica" of Bucharest (UPB), Romania
Shkelzen Cakaj, Telecom of Kosovo / Prishtina University, Kosovo
Enzo Alberto Candreva, DEIS-University of Bologna, Italy
Rodrigo Capobianco Guido, São Paulo State University, Brazil
Hakima Chaouchi, Telecom SudParis, France
Silviu Ciochina, Universitatea Politehnica din Bucuresti, Romania
José Coimbra, Universidade do Algarve, Portugal
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Noel Crespi, Institut TELECOM SudParis-Evry, France
Leonardo Dagui de Oliveira, Escola Politécnica da Universidade de São Paulo, Brazil
Gerard Damm, Alcatel-Lucent, USA
Francescantonio Della Rosa, Tampere University of Technology, Finland
Chérif Diallo, Consultant Sécurité des Systèmes d'Information, France
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD, Germany
Jawad Drissi, Cameron University , USA
António Manuel Duarte Nogueira, University of Aveiro / Institute of Telecommunications, Portugal
Alban Duverdier, CNES (French Space Agency) Paris, France
Nicholas Evans, EURECOM, France
Fabrizio Falchi, ISTI - CNR, Italy
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Robert Forster, Edgemount Solutions, USA
John-Austen Francisco, Rutgers, the State University of New Jersey, USA
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Shauneen Furlong , University of Ottawa, Canada / Liverpool John Moores University, UK
Ana-Belén García-Hernando, Universidad Politécnica de Madrid, Spain
Bezalel Gavish, Southern Methodist University, USA
Christos K. Georgiadis, University of Macedonia, Greece
Mariusz Glabowski, Poznan University of Technology, Poland
Katie Goeman, Hogeschool-Universiteit Brussel, Belgium
Hock Guan Goh, Universiti Tunku Abdul Rahman, Malaysia
Pedro Gonçalves, ESTGA - Universidade de Aveiro, Portugal
Valerie Gouet-Brunet, Conservatoire National des Arts et Métiers (CNAM), Paris
Christos Grecos, University of West of Scotland, UK
Stefanos Gritzalis, University of the Aegean, Greece
William I. Grosky, University of Michigan-Dearborn, USA
Vic Grout, Glyndwr University, UK
Xiang Gui, Massey University, New Zealand
Huaqun Guo, Institute for Infocomm Research, A*STAR, Singapore
Song Guo, University of Aizu, Japan

Ching-Hsien (Robert) Hsu, Chung Hua University, Taiwan
Javier Ibanez-Guzman, Renault S.A., France
Lamiaa Fattouh Ibrahim, King Abdul Aziz University, Saudi Arabia
Theodoros Iliou, University of the Aegean, Greece
Mohsen Jahanshahi, Islamic Azad University, Iran
Antonio Jara, University of Murcia, Spain
Carlos Juiz, Universitat de les Illes Balears, Spain
Adrian Kacso, Universität Siegen, Germany
György Kálmán, ABB AS, Norway
Eleni Kaplani, Technological Educational Institute of Patras, Greece
Behrouz Khoshnevis, University of Toronto, Canada
Ki Hong Kim, ETRI: Electronics and Telecommunications Research Institute, Korea
Atsushi Koike, Seikei University, Japan
Ousmane Kone, UPPA - University of Bordeaux, France
Dragana Krstic, University of Nis, Serbia
Archana Kumar, Delhi Institute of Technology & Management, Haryana, India
Romain Laborde, University Paul Sabatier (Toulouse III), France
Massimiliano Laddomada, Texas A&M University-Texarkana, USA
Thomas D. Lagkas, University of Western Macedonia, Greece
Wen-Hsing Lai, National Kaohsiung First University of Science and Technology, Taiwan
Zhihua Lai, Ranplan Wireless Network Design Ltd., UK
Jong-Hyouk Lee, INRIA, France
Wolfgang Leister, Norsk Regnesentral, Norway
Elizabeth I. Leonard, Naval Research Laboratory - Washington DC, USA
Jia-Chin Lin, National Central University, Taiwan
Chi (Harold) Liu, IBM Research - China, China
Diogo Lobato Acatauassu Nunes, Federal University of Pará, Brazil
Andreas Loeffler, Friedrich-Alexander-University of Erlangen-Nuremberg, Germany
Michael D. Logothetis, University of Patras, Greece
Renata Lopes Rosa, University of São Paulo, Brazil
Hongli Luo, Indiana University Purdue University Fort Wayne, USA
Christian Maciocco, Intel Corporation, USA
Dario Maggiorini, University of Milano, Italy
Maryam Tayefeh Mahmoudi, Research Institute for ICT, Iran
Krešimir Malarić, University of Zagreb, Croatia
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Herwig Mannaert, University of Antwerp, Belgium
Adrian Matei, Orange Romania S.A, part of France Telecom Group, Romania
Natarajan Meghanathan, Jackson State University, USA
Emmanouel T. Michailidis, University of Piraeus, Greece
Ioannis D. Moscholios, University of Peloponnese, Greece
Djafar Mynbaev, City University of New York, USA
Pubudu N. Pathirana, Deakin University, Australia
Christopher Nguyen, Intel Corp., USA
Lim Nguyen, University of Nebraska-Lincoln, USA
Brian Niehöfer, TU Dortmund University, Germany

Serban Georgica Obreja, University Politehnica Bucharest, Romania
Peter Orosz, University of Debrecen, Hungary
Patrik Österberg, Mid Sweden University, Sweden
Harald Øverby, ITEM/NTNU, Norway
Tudor Palade, Technical University of Cluj-Napoca, Romania
Constantin Paleologu, University Politehnica of Bucharest, Romania
Stelios Papaharalabos, National Observatory of Athens, Greece
Gerard Parr, University of Ulster Coleraine, UK
Ling Pei, Finnish Geodetic Institute, Finland
Jun Peng, University of Texas - Pan American, USA
Cathryn Peoples, University of Ulster, UK
Dionysia Petraki, National Technical University of Athens, Greece
Dennis Pfisterer, University of Luebeck, Germany
Timothy Pham, Jet Propulsion Laboratory, California Institute of Technology, USA
Roger Pierre Fabris Hoefel, Federal University of Rio Grande do Sul (UFRGS), Brazil
Przemyslaw Pochec, University of New Brunswick, Canada
Anastasios Politis, Technological & Educational Institute of Serres, Greece
Adrian Popescu, Blekinge Institute of Technology, Sweden
Neeli R. Prasad, Aalborg University, Denmark
Dušan Radović, TES Electronic Solutions, Stuttgart, Germany
Victor Ramos, UAM Iztapalapa, Mexico
Gianluca Reali, Università degli Studi di Perugia, Italy
Eric Renault, Telecom SudParis, France
Leon Reznik, Rochester Institute of Technology, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
David Sánchez Rodríguez, University of Las Palmas de Gran Canaria (ULPGC), Spain
Panagiotis Sarigiannidis, University of Western Macedonia, Greece
Michael Sauer, Corning Incorporated, USA
Marialisa Scatà, University of Catania, Italy
Zary Segall, Chair Professor, Royal Institute of Technology, Sweden
Sergei Semenov, Broadcom, Finland
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Dimitrios Serpanos, University of Patras and ISI/RC Athena, Greece
Adão Silva, University of Aveiro / Institute of Telecommunications, Portugal
Pushpendra Bahadur Singh, MindTree Ltd, India
Mariusz Skrocki, Orange Labs Poland / Telekomunikacja Polska S.A., Poland
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica Spahiu, University of Craiova, Romania
Young-Joo Suh, POSTECH (Pohang University of Science and Technology), Korea
Hailong Sun, Beihang University, China
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Fatma Tansu, Eastern Mediterranean University, Cyprus
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Božo Tomas, HT Mostar, Bosnia and Herzegovina
Piotr Tyczka, Poznan University of Technology, Poland

John Vardakas, University of Patras, Greece
Andreas Veglis, Aristotle University of Thessaloniki, Greece
Luís Veiga, Instituto Superior Técnico / INESC-ID Lisboa, Portugal
Calin Vladeanu, "Politehnica" University of Bucharest, Romania
Natalija Vlajic, York University - Toronto, Canada
Benno Volk, ETH Zurich, Switzerland
Krzysztof Walczak, Poznan University of Economics, Poland
Krzysztof Walkowiak, Wroclaw University of Technology, Poland
Yang Wang, Georgia State University, USA
Yean-Fu Wen, National Taipei University, Taiwan, R.O.C.
Bernd E. Wolfinger, University of Hamburg, Germany
Riaan Wolhuter, Universiteit Stellenbosch University, South Africa
Yulei Wu, Chinese Academy of Sciences, China
Mudasser F. Wyne, National University, USA
Gaoxi Xiao, Nanyang Technological University, Singapore
Bashir Yahya, University of Versailles, France
Abdulrahman Yarali, Murray State University, USA
Mehmet Erkan Yüksel, Istanbul University, Turkey
Pooneh Bagheri Zadeh, Staffordshire University, UK
Giannis Zaoudis, University of Patras, Greece
Liaoyuan Zeng, University of Electronic Science and Technology of China, China
Rong Zhao , Detecon International GmbH, Germany
Zhiwen Zhu, Communications Research Centre, Canada
Martin Zimmermann, University of Applied Sciences Offenburg, Germany
Piotr Zwierzykowski, Poznan University of Technology, Poland

International Journal on Advances in Telecommunications
Volume 8, Numbers 1 & 2, 2015

CONTENTS

Feasibility Study of a PLC System for Avionic Safety-Critical Systems

Thomas Larhzaoui, Fabienne Nouvel, Jean-Yves Baudais
IETR
Rennes, France
thomas.larhzaoui@insa-rennes.fr, fabienne.nouvel@insa-rennes.fr, jean-yves.baudais@insa-rennes.fr

Virginie Degardin, Pierre Laly
IEMN, TELICE
Lille, France
virginie.degardin@univ-lille1.fr, pierre.laly@univ-lille1.fr

***Abstract*— To increase the flexibility of the aircraft equipments and to reduce the possession and operating costs of the aircrafts, the main aircraft manufacturers want to change fluidic systems by electrical systems. However, this evolution induces a high increase of the number of wires. Reducing the amount of wiring also allows decreasing the construction and the maintenance costs, and the polluting emissions. Another interest is improving the reliability of aircraft equipment such as allowing monitoring of power cables. To limit the number wires, we proposed to use power line communication (PLC) for flight control systems (FCS) on the high voltage direct current network (HVDC), between a calculator unit and a power inverter for medium-haul aircrafts. PLC technology has proven its reliability for indoor network with the Homeplug Av standard. Nowadays, many studies deal with the possibility to use PLC for embedded systems. However, PLC for safety-critical avionic systems are not often studied. This paper attempts to define the physical layer for such application. The proposed transmission technique used is orthogonal frequency division multiplexing (OFDM), which is widely used with success in many telecommunication systems. In this paper, we present throughput measurement with Homeplug Av modems to prove the feasibility of PLC in aircraft environment. However, the Homeplug Av parameters are not adapted to the aeronautic constraints. Based on channel transfer function measurements and analysis, we proposed to adapt the OFDM parameters to comply with the FCS real-time constraints.**

Keywords-PLC; OFDM; coherence bandwidth; delay spread; insertion gain; channel impulse response; aircraft; avionic; bit rate; safety critical systems; HVDC network.

I. INTRODUCTION

In future aircrafts, hydraulic flight control systems (FCS) will be replaced by electric ones. The main interests are a better flexibility and a decrease in maintenance costs. However, the major problem is the increasing of wires length. Since the actuators for the FCS are electrical actuators, it is possible to change the medium to improve the speed of the transmissions, the reliability or to decrease the complexity of the electrical network. To reduce the electrical network complexity and propose a reliable transmission in the medium-haul aircrafts, PLC technology seems to be a good candidate. Reducing the mass of wiring also has the benefit of reducing not only the maintenance and the construction costs, but also the polluting emissions. In addition, the PLC technology could improve the reliability of the aircraft equipment through the monitoring of the power wires. In previous work [1], we proposed to define OFDM parameters for such kind of transmission in order to comply with the real-time constraints. It is also possible to use optical fibers for the FCS, which is called fly-by-light [2]. They allow high data rates and fast transmissions. However, it does not solve the problems of the network complexity, and aircraft manufacturers remain reluctant to use them for avionics critical systems due to the maintenance constraints. Another possibility is to use wireless communication which is called fly-by-wireless [3]. The main advantage of this technology is that it removes the wired medium. However, this technology is vulnerable from the electromagnetic point of view, in terms of safety and reliability.

PLC technology has proven its reliability in in-home network with HomePlug Av [4]. This standard allows to transmit data with a bit rate of about 200 Mbit/s in the [1;30] MHz bandwidth. In addition, there are numerous studies concerning PLC in different kinds of vehicles like cars [5][6][7][8], ships [10][11], and trains [12][13]. PLC technology is also investigated in aircraft cabin lighting system for multimedia application in the European project TAUPE [14][15]. However, even if the cabin lighting system network is representative of one part of the aircraft electrical network, it is not appropriate for safety-critical systems like FCS. A first study, which proposed to use PLC for a critical system, has been done in [16]. The authors studied the feasibility of using PLC technology between the power inverter and the actuator for landing gears. In this case, the wire length between the power inverter and the actuator is about five meters and the network is a non-filtered low voltage AC network. Nevertheless, in this paper, we focus on a new high voltage direct current (HVDC) network, which is longer (until thirty two meters long) and filtered.The HVDC network is a new ± 270 VDC power supply network, which will replace the AC 115 V, 400 Hz network.

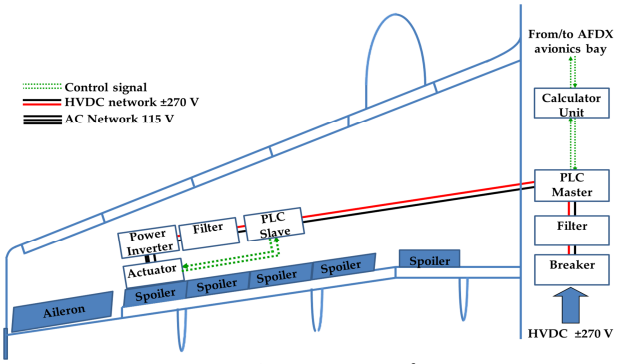


Figure 1. PLC system on aircraft wing

The interest for the aircraft manufacturers to use a HVDC network is to allow the simplification of the power networks (alternators, power conversion, circuit breakers, etc.). It becomes possible to use the reversibility of the electrical actuators for the FCS to produce electrical power. As the HVDC network is still in design, we can influence the design of the power network in order to improve the quality of the propagation channel for the proposed PLC transmissions.

FCS do not require a high bit rate link, few Mbit/s being enough. Nevertheless, the communication must be highly reliable, deterministic, real time and must comply with theDO-160 [17]. The DO-160 specifies test conditions for the design of avionics electronic hardware in airborne systems. As shown in Fig. 1, we consider the link between the calculator unit and the power inverter located near the actuators used for flight control. In this illustration, the PLC master near the calculator unit transmits data to the PLC slave near the power inverter. It corresponds to a point-to-point topology. It is also possible to use point-to-multipoint topology, where one PLC node transmits data to two PLC nodes. Moreover, one of the major challenges of the command of the FCS is the real time constraints. The FCS operates at frequency about 1 kHz, which is called the fast control loop. According to the common practice for the aeronautic equipment, command systems must work six times faster than the equipment that they command. It represents a 6 kHz frequency system or a 167 µs period in this scheme. However, there are several calculators in this fast control loop, which require time processing. It seems reasonable to consider that the PLC system time processing of the must not exceed from 10 % to 20 % of the 167 µs period. In our case, it varies from 17 µs to 34 µs.

The proposed PLC data transmissions are based on the OFDM technique [18], which has been used with success in many wireless and wired line communication systems like DVB, indoor PLC standards or 3GPP-LTE. This technology is interesting for the PLC transmission because it is flexible and robust in frequency selective channels.

In this paper, we measure and analyze the propagation channel in order to define an OFDM symbol duration in compliance with the real-time constraints. Bit rate measurements are also performed to prove the feasibility of the PLC on the HVDC network for the FCS.

This paper is organized as follows. In Section II, we describe the channel and the test bench, while results on the insertion gain are presented in Section III. Section IV describes the channel analysis and the optimization of the OFDM parameters is presented in Section V. In Section VI, simulations are performed in order to check the parameters proposed in the previous section. The bit rate measurements of the PLC link are given in Section VII. A synthesis of the main results and a conclusion are given in Section VIII.

II. DESCRIPTION OF THE TEST BENCH AND OF THE MEASUREMENT CONFIGURATION

In the test bench, the channel is composed of a harness and two couplers that allow to connect the communication system over the HVDC network. Two kinds of couplers are used: capacitive couplers or inductive couplers.

A. *Harnesses Configuration*

During the measurement campaign, three architectures were studied:

- the point-to-point architecture, with two capacitive couplers: architecture 1 (Fig. 2),
- the point-to-point architecture, with two inductive couplers: architecture 2 (Fig. 3),
- the point-to-multipoint architecture with one master and two slaves, architecture 3 (Fig. 4).

The tests have been performed on a test bench with active loads which are representative of actual avionic loads and with a ± 270 DC power supply. A fan has also been used. For this experiment, a harness of 32 meters long is used. It is representative of one possible wire length of the power network of the FCS in aircrafts. It includes one twisted pair, one twisted quadrifilar cable, and one single wire.

In Fig. 2, the architecture 1 is represented. Capacitive couplers are used to transmit on one twisted pair over +270 V and -270 V. Capacitive couplers are composed by a transformer for the galvanic isolation and two capacitors. The harness is composed by one twisted pair. We have also considered another possibility, i.e., to use a twisted quadrifilar for two different transmissions on the two polarities for the same load. In this case, the quadrifilar is short circuited at both ends of each polarity and inductive couplers can be used as illustrated in the architecture 2 in Fig. 3. In this figure, one signal is transmitted on the pair on the + 270 V and one other signal is transmitted on the other pair on the - 270 V. It must be emphasized that, for the same DC power, the diameter of each wire of the quadrifilar can be reduced to avoid an increase of the copper weight.

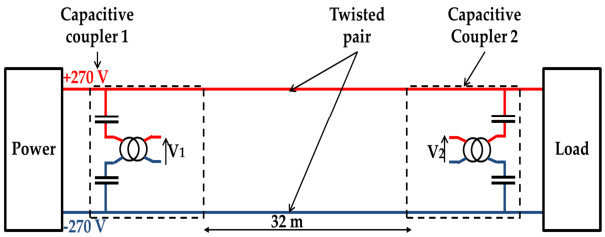


Figure 2. Point-to-point architecture with capacitive coupler

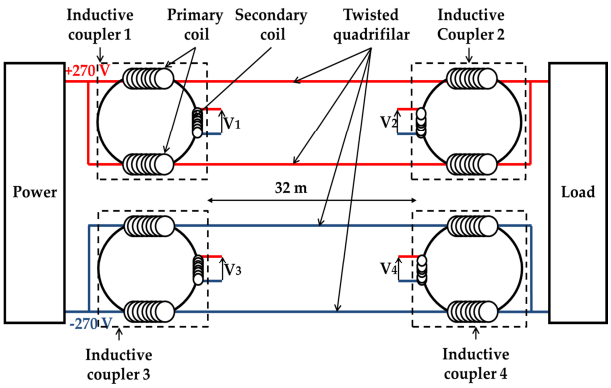


Figure 3. Point-to-point architecture with inductive coupler

There are two main interests to use this architecture. First, due to the short circuits at both ends, the communications are less disturbed by the noise produced by the power supply and the load. The other one is the possibility to use the avionics full duplex (AFDX) twisted quadrifilar already used in aircrafts for the transmissions [19]. These quadrifilars are light and their conception is mature. It means that the physical characteristics of the wires and the connectors are well known, which is the major asset for the implantation in the aircrafts. In addition, this architecture allows different possibilities of transmission. For example, it is possible to use the second polarity for redundancy or use the four couplers for a full duplex transmission.

The architecture 3 is presented in Fig. 4. In this case, three couplers are used on the +270 V. Thus, the harness is composed of one quadrifilar and one wire for the -270 V polarity. Such architecture allows to test the case where one effector as an aileron is driven by two actuators (loads). For the architecture 3, only the transmission on the +270 V polarity has been tested. It is necessary to test the point-to-multipoint topology because the transfer function cannot be deduced from the point-to-point architecture due to the multipath and crosstalk. But, similarly to the architecture 2, it is possible to consider a transmission on both polarities for two loads for the explorations of different possibilities of transmissions.

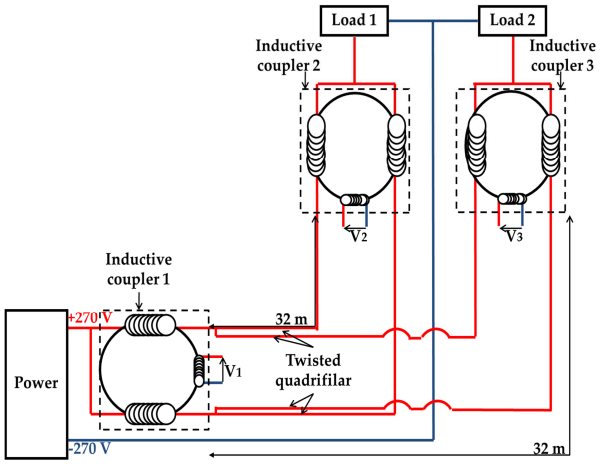


Figure 4. Point-to-multipoint architecture with inductive coupler

B. *Channel Measurements*

Transfer function measurements have been carried out with a network analyser in the [1;100] MHz bandwidth with a 5 kHz resolution bandwidth. To do this, we used the frequency scanning method. This technique involves scanning the channel using a network analyser with a constant step Δf on a frequency bandwidth equal to $f_0+N\Delta f$, where f_0 is the minimum frequency, Δf is the frequency step, and N is the number of measurement points. For each configuration, the transfer functions have been measured between the input and the output V_1, V_2, V_3 and V_4. Since there are more than two couplers on the architectures 2 and 3, 50 Ω loads are connected on the non used couplers during the transfer function measurements.

III. PRELIMINARY RESULTS

In order to prove the PLC feasibility and measure the throughput on the test bench, Homeplug Av modems have been plugged on the architectures 1 and 2 through the couplers. The power supply of the modems comes from the HVDC network via a DC/DC converter. Data transmissions between couplers and modems are done by a twisted pair. In this preliminary study, we just focus on the experimental aspects to obtain a first result on link capabilities using product on the shelf.

The throughput measurements are shown in Fig. 5 and have been performed with the Jperf software. When the network is turned off, the throughput achieved 40 Mbit/s for both architectures. The tests with the active loads show a small decrease of the throughput. The tests with the fan, which is noisy, show the interest of the inductive coupler compared with the capacitive coupler. Indeed, the throughput does not change a lot with the inductive coupler but is divided by two with the capacitive coupler.

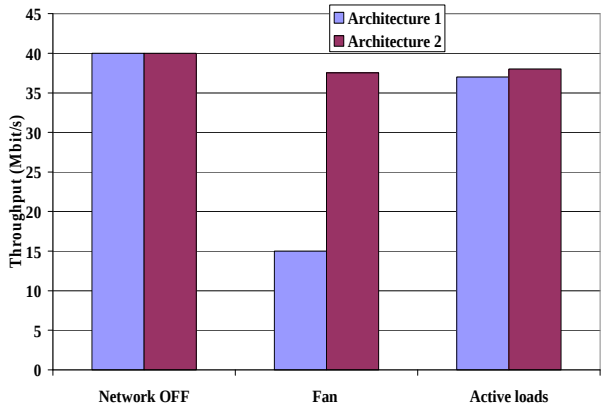


Figure 5. Throughput measurement

The minimum bit rate measured is about 15 Mbit/s. As specified in the communication buses like ARINC 429 [20], MIL-STD-1553 [21] and ARINC 825 [22], the required throughput for the FCS is about 1 Mbit/s. Thus, the results presented in Fig. 5 confirm the feasibility of the PLC technology in such application and thus it is possible to use PLC technology to convey such buses for FCS.

IV. INSERTION GAIN AND CHANNEL IMPULSE RESPONSE

In Section III, throughput is measured for the architectures 1 and 2. Even if this throughput is sufficient for FCS, the Homeplug Av standard is not adapted for the aircraft communication. In fact, the OFDM symbol duration is equal to 46.52 µs, which is not in accordance with the real time constraints. Thus, we measured the channel transfer function to define new physical layer parameters adapted to this channel for a safety-critical aeronautical system.

A. Tested configurations

In Table I, the tested configurations are presented. "OFF" means that the power supply and the loads are connected to the network and they are turned off. "ON" means that the power supply and the loads are turned on. Transfer functions are measured between the two couplers. "P-to-p" means point-to-point and "P-to-m" means point-to-multipoint.

B. Channel measurement results

In this paragraph, we only show the transfer function with the network "ON" because there are no major differences between the transfer function with the network "OFF". In addition, we do not represent the transfer function on the-270 V with the architecture 2 because, due to the symmetry of the network, transfer functions are similar on both polarities.

Fig. 6 represents the insertion gain (IG) for the point-to-point topology, namely, the architectures 1 and 2. For the architecture 1, which corresponds to the configurations C2 and C4, the IG decreases over the entire bandwidth with several resonances. For the architecture 2, which corresponds to the configuration C8, the IG first decreases linearly (in dB) with the frequency up to 40 MHz, and varies from -5 to -25 dB. Then, the IG remains nearly constant between 40 MHz and 80 MHz and, beyond 80 MHz, decreases very rapidly. Fig. 7 shows the cumulative distribution function (CDF) of the IG for the architectures 1 and 2. For the configuration C2 the insertion gain is higher than -23 dB over 50 % of the bandwidth and higher than -32 dB over 90 % of the entire bandwidth. For the configuration C4 the insertion gain is higher than -28 dB over 50 % of the bandwidth and higher than -35 dB over 90 % of the entire bandwidth

TABLE I. TESTED CONFIGURATIONS

	Configuration	Topology	Coupler	Polarity	Power	Loads
Architecture 1	C1	P-to-p	Capacitive	±	OFF	Active loads
	C2				ON	
	C3				OFF	Fan
	C4				ON	
Architecture 2	C5		Inductive	-	OFF	
	C6			+	OFF	
	C7			-	ON	
	C8			+	ON	
Architecture 3	C9	P-to-m			OFF	Active loads
	C10				OFF	Fan
	C11				ON	Active loads
	C12				ON	Fan

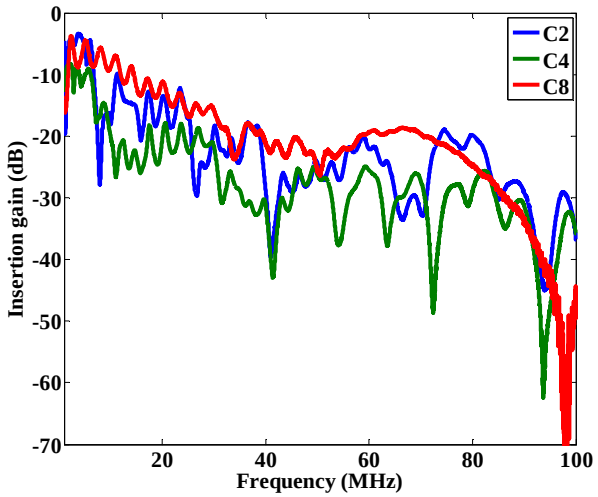


Figure 6. Insertion gain for the architectures 1 and 2

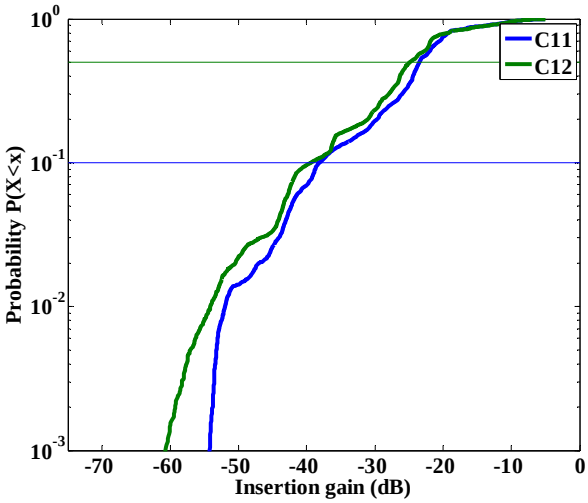


Figure 9. CDF the architecture 3

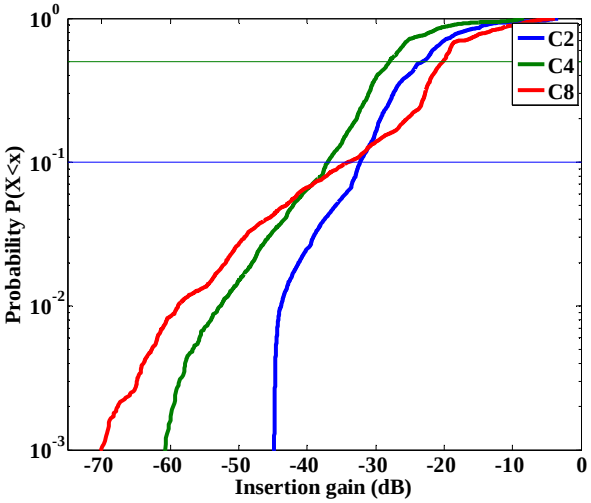


Figure 7. CDF the architectures 1 and 2

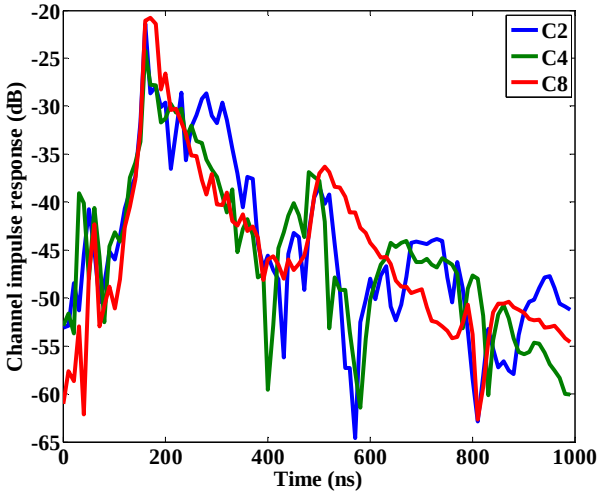


Figure 10. Channel impulse for architecture 1 and architecture 2 in the [1;100] MHz bandwidth

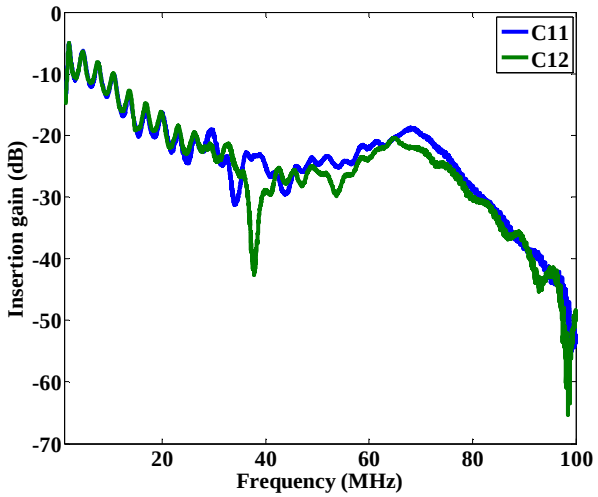


Figure 8. Insertion gain for the architecture 3

Fig. 8 represents the IG for the point-to-multipoint topology, namely, the architecture 3. The IG is quite similar for the two configurations. There is a resonance at about 39 MHz for configuration C12. The IG of the architecture 3 is few dB lower than the IG of the architecture 2. Finally, Fig. 9 shows the CDF for the architecture 3. For the configuration C2 the insertion gain is higher than -25 dB over 50 % of the bandwidth and higher than -39 dB over 90 % of the entire bandwidth.

The channel may be studied also in the time domain in order to get the impulse response The channel impulse response has been obtained from the measurements of the complex transfer function by applying a 20000 points inverse Fourier transform (IFFT) in the [1;100] MHz bandwidth. The results of the channel responses for the architectures 1 and 2 are shown in Fig. 10.

Finally, the coherence bandwidth is calculated from the transfer function and the delay spread is calculated from the channel impulse response. The coherence bandwidth and the delay spread allow to define the subcarrier spacing and the cyclic prefix duration, respectively.

V. COHERENCE BANDWIDTH AND DELAY SPREAD

The coherence bandwidth is deduced from the absolute value of the autocorrelation of the complex transfer function [23]. The values of both coherence bandwidth and delay spread are calculated for 8 different frequency bandwidths for each configuration, from the [0;20] MHz bandwidth to the [0;100] MHz bandwidth with a step of 10 MHz. In the following, the coherence bandwidth is calculated for a correlation coefficient of 0.9. The delay spread is calculated from the channel impulse responses according to [24]. It appears that the values of the values of the coherence bandwidth and delay spread are quite independent of the frequency bandwidth analysis. Fig. 11 represents the coherence bandwidth versus the inverse of the delay spread for all the cases (bandwidths and configurations). Conventionally, the coherence bandwidth is proportional to the inverse of the delay spread. In our case, the linear regression leads to a correlation coefficient of 0.75. This result has been also noticed for indoor networks [25].

It appears that the architectures and the kind of coupler do not have a strong impact on the channel characteristics, the coherence bandwidth being of the order 700–1200 kHz, as the delay spread varies from 60 to 110 ns. These results are quite similar to those obtained for other embedded systems as shown in Table II. The delay spread measured for these channels is between 34 ns and 380 ns and the coherence bandwidth is between 0.4 MHz and 0.9 MHz.

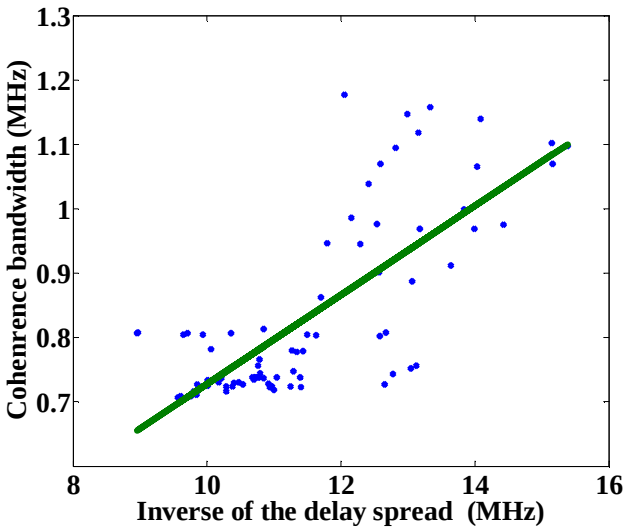


Figure 11. Coherence bandwidth versus of the inverse of delay spread

TABLE II. COHERENCE BANDWIDTH AND DELAY SPREAD FOR DIFFERENT VEHICLES

Vehicles References	Bandwidth (MHz)	Delay spread (ns)	Coherence bandwidth (MHz)
Car [15]	[1;50]	34-200	0.4-4.8
Aircraft [16]	[1;30]	100	0.6-0.9
Car[25]	[0.3;100]	130	0.48
Car [26]	[1;70]	380	0.4-0.7

VI. OPTIMIZATION OF OFDM PARAMETERS

Taking into account the obtained results, the next step is to adapt the OFDM symbol duration to the real time constraint. The real time constraint is defined by the duration between the moment when a bit enter in the transmitter and the moment when the same is available at the output of the receiver. We assume that all the information contained in an OFDM symbol must be completely received to be considered usable. As a result, the duration of the OFDM symbol is considered as an incompressible latency time. Thus, it is necessary to ensure that the OFDM symbol duration is lower than the real time constraints. It leads to an OFDM symbol duration between 17 µs and 34 µs, as explained in Section I. The OFDM symbol duration T_{OFDM} is given by the equation:

$$T_{OFDM} = \frac{1}{sub\,carrier\,spacing} + cyclic\,prefix\,duration \quad (1)$$

Thus, we need to adapt the sub-carrier spacing and the cyclic prefix duration (CP).

A. OFDM Sub-carrier Spacing

In order to meet the real time constraints, it is necessary to minimize the processing time for in the physical layer. Since fast Fourier transform (FFT) is a time consuming process proportional to the number of sub-carriers, one can try to decrease the number of carriers and choose, as in common practice, a sub-carrier spacing less than 10 % of the coherence bandwidth. Taking into account the values in Fig. 10, this leads to a 70 kHz sub-carrier spacing, which is about three times the value given in Homeplug Av specifications (24.414 kHz). To decrease the time processing, it is better to use a FFT size of power of 2. Finally, it is possible to switch off sub-carriers to transmit data on the proper frequency bandwidth. In our case, it leads to 428 or 1428 useful sub-carriers for a transmission bandwidth over 30 MHz or 100 MHz respectively.

B. Interference Characterization

Using the channel impulse response values, it is also possible to compute the inter symbol interference (ISI) and the inter carrier interference (ICI) according to the cyclic prefix CP length.

Then, it becomes possible to choose the optimal CP length because the increase of the CP length decreases the power spectral density of ISI and ICI but also reduces spectral efficiency and data rate. The power spectral density of ISI and ICI can be computed by the equation [27]:

$$I_{ISI+ICI}(n) = 2\sigma_x^2 \sum_{l=L_{cp}+1}^{L_c-1} \left| \sum_{u=l}^{L_c-1} h(u) \exp\left(-j\frac{2\pi}{N} un\right) \right|^2 \quad (2)$$

where σ_x^2 is the variance of modulated signal, h is the channel impulse response, L_c the channel length expressed in number of samples, L_{cp} being also expressed in terms of number of samples, N the number of sub-carriers, and n the frequency sample index.

Fig. 12 gives the normalized PSD interferences, expressed in dBm/Hz for configuration C11, which presents the highest delay spread, and calculated in the [1;30] MHz bandwidth. The PSD of the interferences has been plotted versus the sub-carrier index and for various lengths of the CP. As expected, the interference PSD decreases rapidly with the length of the cyclic prefix but, beyond 20 samples, it does not vary appreciably. Thus, it is not necessary to use a L_{cp} value higher than 20 samples.

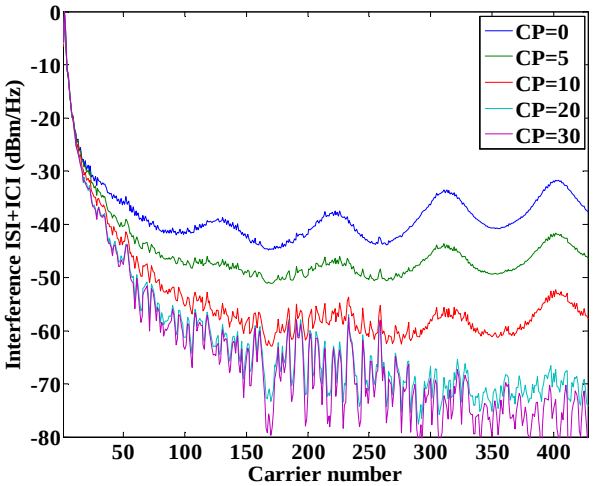


Figure 12. Interference in the [1-30] MHz bandwidth

C. Simulation Results

To show the influence of the CP length on the bit error rate (BER), an OFDM transmission chain has been simulated using Matlab. The simulated transmission chain is presented in Fig. 13. The first block of the transmitter is a random binary data generator. The generated data are mapped using binary phase shift keying (BPSK) modulation and an IFFT is then applied. The CP is then added to the OFDM symbol in the time domain. The channel includes both the complex channel impulse response and the additive withe Gaussian noise (AWGN). At the receiver, the inverse process is realized and an equalization is used to compensate the distortion effect introduced by the channel. For theses simulations, the channel estimation is assumed to be ideal and the zero forcing equalization is applied [28].

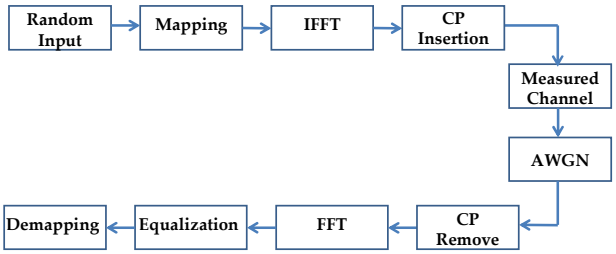


Figure 13. simulated OFDM transmission and reception chain

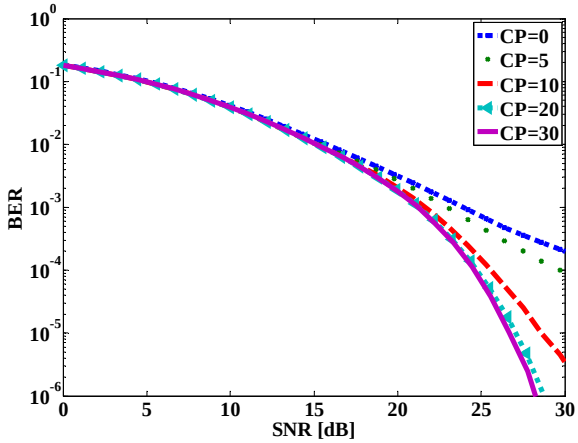


Figure 14. BER for different cyclic prefix length

The time and frequency synchronization are assumed to be perfect. Each useful sub-carrier transmits one bit corresponding to a BPSK symbol. In the following simulations, the configuration C11 is tested over the [0;36] MHz bandwidth and all the sub-carriers in the [30;36] MHz bandwidth are turned off. The SNR is defined as the ratio between the received power and the noise power. Fig. 14 shows the influence of the CP length (in number of samples) on the BER. In this configuration, the sub-carrier spacing is equal to 70 kHz. The degradation of the BER increases when the CP length decreases. One can observe that a CP length of 20 samples allows to absorb the interference due to the multipaths, as expressed in the Section VI.*B*. Taking into account the interferences and the BER calculations, we propose a CP duration of 20 samples, which correspond to 666 ns on the [1;30] MHz bandwidth. As a comparison, if the cyclic duration was chosen equal to 2 to 4 times the delay spread, as suggested in [29], we would obtain a CP duration between 220 and 440 ns, which is not sufficient in our case as observed in Fig.14.

VII. Synthesis and Conclusion

In this paper, we have studied the feasibility of PLC transmissions for avionic safety critical systems. Throughput measurements with Homeplug Av modems have been done and show a sufficient throughput for the FCS. In addition, the channel measurements prove that it is possible to reduce the duration of an OFDM symbol, compared to the Homeplug Av standard, by both increasing the sub-carrier spacing and decreasing the cyclic prefix

duration. In fact, in the Homeplug Av standard, the sub-carrier spacing is 24.414 kHz, the minimum cyclic prefix duration is 5.56 µs, and the OFDM symbol duration is 46.52 µs. In this PLC application, we propose to increase the sub-carrier spacing up to 70 kHz and to decrease the CP duration to 666 ns. Consequently, the OFDM symbol duration is 14.94 µs. These results will help us to define the physical layer parameters for a PLC avionics system in accordance with real-time constraints of a fast control loop. This study can be applied to other critical avionic systems running on a HVDC network like landing gear. It is also possible to use this study for a slow control loop on HVDC network like thrust reversal.

In the next steps, we will continue to define the OFDM parameters (constellation and frequency bandwidth) and the channel coding to ensure a sufficient quality of the service for the FCS. In addition of the real time constraint, the quality of service that is defined by the useful bitrate (10 Mbit/s), the bit error rate (10^{-12} as on the AFDX), and the respect of the DO-160 gauge in conducted emissions may be taken into account in the parameters of the physical layer.

ACKNOWLEDGMENT

This study is funded by Safran and followed by IETR of Rennes.

REFERENCES

[1] T. Larhzaoui, F. Nouvel, J.Y. Baudais, V. Déguardin and P. Laly, “Analysis of PLC channels in aircraft environment and optimization of some OFDM parameters,” International Conference on Systems and Networks Communications, pp. 65-69, October 2013.

[2] A. Garg, R.I Linda and T. Chowdhury, “Application of fiber optics in aircraft control system & its development,”*International Conference onElectronics and Communication Systems (ICECS)*, pp. 1-5, February 2014.

[3] D. Dinh-Khanh, A. Mifdaoui and T. Gayraud, “Fly-By-Wireless for next generation aircraft: Challenges and potential solutions,”*Wireless Days (WD)*, pp. 1-8, 21-23 November 2012.

[4] HomePlug Av specification, Version 1.1, May 2007.

[5] J. Granado, A. Torralba and J. Chavez, “Using broadband power line communications in non-conventional applications,” IEEE Transactions on consumer electronics, vol. 57, no. 3, pp. 1092-1098, August 2011.

[6] P. Tanguy and F. Nouvel, “In vehicule PLC simulator based on channel measurements,”Intelligent Transport Systems Telecommunications, pp. 9-11, November 2010.

[7] G. Sung, C. Huang and C. Wang, “A PLC transceiver design of in-vehicle power line in FlexRay-based automotive communication systems,” IEEE International Conference on Consumer Electronics, vol. 13, no. 16, pp. 309-310, January 2012.

[8] M. Mohammadi, L. Lampe, M. Lok, S. Mirabbasi, M. Mirvakili, R. Rosales and P. van Veen, “Measurement study and transmission for in-vehicle power line communication,” IEEE International Symposium on Power Line Communications and Its Applications, pp. 73-78, March 2009.

[9] S. Barmada, L. Bellanti, M. Raugi, and M. Tucci, “Analysis of power-line communication channels in ships,” IEEE Transactions on Vehicular Technology, vol. 59, no. 7, pp. 3161-3170, September 2010.

[10] M. Antoniali, A. Tonello, M. Lenardon and A. Qualizza, “Measurements and analysis of PLC channels in a cruise ship,”IEEE International Symposium on Power Line Communications and Its Applications, vol. 3, no. 6, pp.102-107, April 2011.

[11] A. Akinnikawe and K. Butler-Purry, “Investigation of broadband over power line channel capacity of shipboard power system cables for ship communication networks,”IEEE Power & Energy Society General Meeting, pp.1-9, July 2009.

[12] S. Barmada, A. Gaggelli, A. Musolino, R. Rizzo, M. Raugi and M. Tucci, “Design of a PLC system onboard trains: Selection and analysis of the PLC channel,”IEEE International Symposium on Power Line Communications and Its Applications, vol. 2, no. 4, pp. 13-17, April 2008.

[13] K. Liu D. Jiang, “PLC used in the train control simulation system,”IEEE International Conference on Computer Science and Automation Engineering, vol. 2, no. 25-27, pp. 308-311, May 2012.

[14] S. Bertuol, I. Junqua, V. Degardin, P. Degauque, M. Lienard, M. Dunand and J. Genoulaz , “Numerical assessment of propagation channel characteristics for future application of power line communication in aircraft,”EMC Europe, pp. 506-511, September 2011.

[15] V. Degardin, I. Junqua, M. Lienard, P. Degauque and S. Bertuol, “Theoretical Approach to the Feasibility of Power-Line Communication in Aircrafts,”IEEE Transactions on Vehicular Technology, vol. 62, no. 3, pp. 1362-1366, March 2013.

[16] K. Kilani, V. Degardin, P. Laly, M. Lienard and P. Degauque, “Impulsive noise generated by a pulse width modulation inverter: modeling and impact on powerline communication,” IEEE International Symposium on Power Line Communications and Its Applications, pp. 75-79, March 2013.

[17] DO-160, “Environmental conditions and test procedures for airborne equipment,” RTCA Inc., 2007.

[18] K. Fazel and S. Kaiser, *Multi-Carrier and Spread Spectrum Systems: From OFDM and MC-CDMA to LTE and WiMAX*, John Wiley \& Sons Ltd, 2008.

[19] ARINC 664, “ Aircraft data network part 7 avionics full duplex switched ethernet (AFDX) network,” Arinc specification 664 P, Jun 2005.

[20] Arinc 429, “Specification Tutorial”, AIM GmbH, July 2001.

[21] “An Interpretation of MIL-STD-1553B,” SBS Technologies Inc.

[22] R. Knueppel, “Standardisation of CAN network for airborne use through ARINC 825,” International CAN Conference, pp.1-8, March 2012.

[23] T.S. Rappaport, *Wireless communication principle and practice*, Prentice Hall, 1996.

[24] M. Tlich, G. Avril and A. Zeddam, “Coherence bandwidth and its relationship with the rms delay spread for PLC channels using measurements up to 100 MHz,” Home networking, International Federation for Information Processing, vol. 256, pp. 129-142, January 2008.

[25] A.B. Vallejo-Mora, J.J. Sanchez-Martinez, F.J. Canete, J.A. Cortés and L. Diez, “Characterization and evaluation of in-vehicle power line channels,”IEEE Global Telecommunications Conference, pp. 1-5, December 2010.

[26] M. Lienard, M. Olivas Carrion, V. Degardin, and P. Degauque, “Modeling and analysis of in-vehicle power line communication Channels,”IEEE Transaction on Vehicular Technology, vol. 57, no. 2, pp. 670-679, March 2008.

[27] W. Henkel, G. Taubock, P. Odling, P.O. Borjesson and N. Petersson, “The cyclic prefix of OFDM/DMT. An analysis,”International Zurich Seminar on Broadband Communications, Access, Transmission, Networking, pp. 22-1-22-3, September 2002.

[28] G. Proakis, *Digital communication*, McGraw Hill series in electrical and computer engineering, 1996.

[29] R. Van Nee and R. Prasad, *OFDM for wireless multimedia communications*, Artech House, 2000.

Musing: A Mobile Client and Web Server Augmented Reality Application for Museum Visitors and Curators

K. Whiteside[1], G. Atkinson[1], M. Stump[2], G. Lawrence[2], D. E. Tamir[1]

[1]Department of Computer Science
[2]School of Art and Design
Texas State University
San Marcos, TX, USA
{kjw52, gma23,mr14,gl16,dt19}@txstatet.edu

***Abstract*—Textual didactics, used in museums and galleries provide access to historical, socio-political, technical, and biographic information about the artworks and artists. These types of didactics are considered to be cost-effective. However, they do not enable the use of audio, video, and Web interface that allows for multiple forms of usage for the museum visitors. We have developed a smartphone application, called *Musing*, for interaction of museum visitors with informational content and enhancement of their museum experience. *Musing* is an augmented reality (AR) application that enables the visitor to capture an artwork with a smartphone camera. Using image processing, the application recognizes the artwork and places graphical user interface objects in the form of Points of Interest (POIs) onto the image of the artwork displayed on-screen. These POIs provide the visitor with additional didactic information in the form of text overlays, audio, video, and Web sites. The *Musing* application and administrator Web site, described in this paper, is designed with several performance and efficiency goals, including high reliability and recognition rate, high usability, and significant flexibility. The application is designed to be adaptable to a variety of museums and galleries without requiring special hardware or software. Furthermore, the administrative interface enables museum staff to provide content for the didactics without requiring software development skills.**

Keywords—interactive didactic; museum didactic; virtual museum; image recognition; augmented reality

I. INTRODUCTION

Museums have historically been tasked with providing access to, and educating visitors about artworks. Museum didactics attempt to clarify artworks' meanings by addressing concepts of art, history, politics, construction techniques, as well as the lives of artists. For many visitors, however, museum and gallery exhibitions may lack the proper context to allow access points for exhibited works and can leave the "uninitiated viewer" intimidated, "particularly when it comes to interpretation" [1][2].

In many ways, mobile technologies, such as responsive Web sites and AR, present an ideal opportunity to make those personal connections with the visitor, as well as help the visitor make connections to the exhibited objects and/or works of art. As such, the context for the artwork is broadened via interviews, videos, Web sites, source material, art historical influences, and other artworks with shared conceptual frameworks, all of which can be integrated into a mobile application for the museum. Such a personalization of experience through narrative is a highly effective way to expand the context for the work and deepen viewers' connections as they process and integrate the information into their existing world-view [3].

Nevertheless, under the current paradigm, in order to add audio and video to exhibits, museums must rely on proprietary hardware and software. The hardware must be provided by the institution at significant cost both in capital investment and in maintenance. The software used on these devices is often proprietary for the exhibition, reliant on external hardware installed in the gallery, and must be reprogrammed for new exhibitions. While large museums have the resources to purchase and maintain these systems, small community-based museums often do not.

Pedagogical shifts away from passive museum participation to active participation are occurring in higher education, as well as in museological practices, and reflect the changing needs of the visitor [4]. An enriched learning environment requires incorporating diverse learning styles, which include visual/print, visual/picture, auditory, kinesthetic, and verbal/kinesthetic modalities [4].

A. Problem Statement

In order for museums and galleries to fully meet the needs of their visitors, they must incorporate didactic information that embraces diverse learning styles and present multiple types of didactic information.

An interactive didactic system should be designed to reach the highest number of museums and their visitors, which does not rely on proprietary hardware, the installation of external devices in the gallery, or the need to reprogram the system when exhibits are modified or added.

In order to create a system that does not require proprietary hardware, the system should be developed on mobile hardware that many of the museum visitors already possess. This hardware would include classes of smartphones and tablets running on iOS or Android operating systems.

To minimize the technical burden on institutions, the system should not rely on extra hardware such as Bluetooth or Near Field Communication (NFC) devices.

An administrator panel should be designed to facilitate ease of editing—addition and deletion of content in such a way as to give museologists these abilities without the requirement of software development skills.

Finally, image processing and image recognition algorithms should be used in order to provide the opportunity for the viewers to deepen their connections to artworks by scanning artworks directly, removing the need for external tokens such as Quick Response codes (QR) or number codes to be entered by users.

B. Hypothesis

By using a combination of off-the-shelf image recognition algorithms and unmodified consumer-level hardware, the research team will be able to create a client application that is fast and accurate enough to be usable in a museum, without the need for proprietary hardware, or external tokens. In addition, retrieving exhibition data via a database will allow for a client program that is sufficiently flexible and does not require reprogramming when exhibitions are added or modified.

The proposed interactive didactic system will be designed with a client-server architecture. A database, administered by a Web site, will provide the client application with access to didactic information without the need to permanently store that information on the device. The client application will be programmed for current popular hardware such as a smartphone or a tablet, either owned by the museum visitor or provided in the form of loaners.

Providing museologists with an efficient and usable software tool that facilitates generating new AR exhibitions and editing / modifying existing AR exhibitions (i.e., editing the Musing server) without requiring software development skills will enable widespread usage of the client part of *Musing*.

In order to test the relative success of the application and its acceptance by museum visitors, *Musing* will be deployed in three exhibitions at The University Galleries at Texas State University, a three thousand square foot, university-based, contemporary art exhibition venue. Benchmark testing of the application will be conducted in order to determine recognition accuracy rate and speed. Post-exhibit, exit questionnaires will be given to visitors in order to determine their acceptance of the client application and perceptions of system performance and usability.

C. Proposed Solution

The research team has developed *Musing*, a mobile, image recognition and AR application that runs on consumer-based iOS systems, requires no external tokens or hardware, and does not require reprogramming between exhibits. The application has passed the Apple approval process and is available at [5].

The main contributions of this research is the design, development, and deployment of an end-to-end reliable, usable, and effective AR system that provides a museum visitor with virtual information and provides museum staff with adaptable, cost effective, and easy to maintain virtual museum utility. To date and to the best of our knowledge, this is the only fully functional system that integrates custom-hardware agnostic and custom-software agnostic virtual museum content delivery and administrative support, which does not require hardware to be installed in the exhibition space, and is freely available to consumers.

This paper, which is an expanded version of [1], is organized in the following way: Section II provides background in the form of relevant past research performed by this team, with Section III containing a Literature review. The application deployment of the *Musing* client application, as well as its associated administration back-end is outlined in Section IV, followed by deployment results showcased in Section V. Section VI explains the evaluation of results from both benchmark testing and exit questionnaires given to the museum visitors. Lastly, Section VII outlines the conclusions and future research objectives for *Musing*.

II. BACKGROUND

A. Previous Research

In 2012, the research team developed a series of responsive Web pages triggered by QR codes used in an exhibition at The University Galleries at the Texas State University [6].

In this pilot program, QR codes were included in the tombstone wall labels placed next to artworks in the gallery. These codes, when scanned with reader software on the user's smartphone, presented the visitor with a custom-built Web responsive page for each artwork (Figure 1). These pages provided supplemental didactic information via news articles that pertained to the artwork's subject matter, full artist biographies, video interviews with the artist, photos of the artist's workspace, and links to external Web sites.

Figure 1. Example QR-triggered Web page with artist interview (http://www.txstgalleries.org/michael-henderson)

During the pilot exhibition, the gallery Web site recorded 23 unique visitors per day with an average time on-page of 3 minutes and 37 seconds. The Web pages that were only accessible by the QR codes were responsible for 16 of the 23 unique daily visitors (69%) and the majority of the time on-page (3 minutes and 33 seconds). For comparison, exhibits installed after the pilot test did not include QR codes. The subsequent exhibit showed a decline in both the number of online visitors (-26%) and the amount of time visitors spent on the gallery Web site (-42.5%). This data indicates that when QR codes and their associated didactic information are

included with the artworks in the gallery, there is an increase in online interaction with the visitor.

The experiment with QR codes in the gallery indicated that visitors would use interactive technologies in the gallery and that they would spend the time necessary to consume the extra content. However, a major drawback of the QR codes was the inability for the museum professional to contextually place information within an artwork's representation. The newer AR technology would allow the administrator to place content exactly where it would be most pertinent to the visitor's view of the artwork. For example, a POI could be placed over a specific person or place in an artwork to provide information about the historical or social significance. Lastly, QR code reader software is not created specifically for the needs of museums and galleries, as they are designed to work for a wide variety of applications, from advertising to stock keeping.

Following the positive response to the QR code project, it was decided that the next step in the research should be to create an AR system allowing for information placement within an artwork, and which could be designed specifically for the needs of museums and galleries.

Several aspects of the *Musing* system, such as the pedagogical and art design characteristics are covered in [7][8]. The current paper, as well as [1], concentrate on the user experience and the technological innovation which enables this experience.

III. LITERATURE REVIEW

The literature review is addressing two areas of interest: 1) The relevance and potential for a positive influence of technology in an exhibition setting on the visitor experience, and 2) current use and applications of AR and mobile applications within an exhibition setting.

In his article, "Designing Mobile Digital Experiences," Tallon talks about the "potential of digital technology" as it surpasses its own hype to become a source of enrichment for visitors' learning [9]. This positions the visitor as a collaborator in the process of making meaning by gathering information and connecting them through their personal frame of reference.

Stephen Weil, author of *Making Museums Matter*, advocates for museums to "be more than merely a communicator or a stimulant" [10]. Moving from the traditional (and outmoded) linear model of communication that provides didactic information in an institutional voice via wall labels and gallery talks, to a circular model that promotes—by incorporating technology into the exhibition materials—an enriched environment in which the visitors can partner in the making of meaning by aggregating a variety of information types as well as voices within the information dissemination. Learning environments that qualify as enriched are reflective of a variety of learning modalities: visual, auditory, and kinesthetic [11], which are comprised of seeing, hearing, and interaction. This is imperative if museums visitors are to move toward relating to art in a non-linear manner.

Another influential theory can be found in John Falk's book, "Identity and the Museum Visitor Experience" [12], wherein Falk identifies five key types of visitors who attend museums while also defining visitor motivation. These five key user types fall within the definitions of human need, rather than that of demographics and are characterized in the following ways relative to basic human needs. They are: 1) Explorers–motivated by personal curiosity (i.e., browsers); 2) Facilitators–motivated by other people and their needs (i.e., a parent bringing a child); 3) Experience-Seekers–motivated by the desire to see and experience a place (i.e., tourists); 4) Professional/Hobbyists–motivated by specific knowledge-related goals (i.e., a scholar researching a specific topic); 5) Rechargers–motivated by a desire for a contemplative or restorative experience.

It is through this research and literature review that the research team gained a clearer picture as to the need for, as well as potential ways to make connections and meaning, in assessing audiences based on their desired experience rather than outmoded demographic considerations. As such, learning typologies, alongside Falk's research on the five types of user experiences seen in museums, provides an emerging picture of the important role that technology can play in facilitating a variety of learning styles, as well as, the diversity of user types are found in exhibition settings.

Addressing the second area of interest, a review of existing literature showed a number of teams researching the possibility of using AR to augment the information provided by museum didactics. In most of the cases, however, these didactics rely on proprietary hardware, require reprogramming between exhibitions, or installation of external tokens (e.g., Bluetooth, RFID, and QR) within the museum space. Some work has been done with respect to the challenges of image recognition, but little attention has been paid with regard to integrating custom-hardware/software agnostic image based picture recognition with content delivery.

Bimber et al. have developed a mobile system, named, *PhoneGuide* allowing museum visitors to use mobile phones to detect artworks in a physical museum space [13]. Their method includes image recognition, using the phone's camera, as well as pervasive tracking techniques using a grid of Bluetooth emitters distributed in the space [13]. The reliance on external tokens (e.g., Bluetooth) to assist in the object recognition would require the museum to install new hardware and provide for updates in each gallery space.

Hatala et al. describe a prototype system, called Ec(h)o, developed to provide "spatialized soundscapes" for museum visitors [14]. That is, specialized audio is played for the listeners depending on their position within the museum. The supplied audio is meant to augment the overall experience of the exhibit rather than providing information about artwork.

Jing et al. have developed a mobile augmented reality prototype system which uses image recognition running on specialized hardware to provide additional information on

physical images displayed in museums for Personal Museum Tour Guide Applications [15]. The system uses the SIFT recognition algorithm that employs "coarse to fine" recognition to improve the speed of the process [16]. Nevertheless, some users complained of slow processing speed.

Blockner et al. developed a prototype system which allows users to create virtual museum tours on a mobile app. The mobile device uses NFC to transmit these tours to projectors positioned within the gallery which display the desired information [17].

Miyashitat et al. have developed an interactive device at the Dai Nippon Printing (DNP) Museum Lab at the Louvre Museum (Paris) for use with an exhibition on Islamic Art. This device used a neural network based system to map content of exhibits and was able to recognize three dimensional objects from a single viewpoint, but also relies on purpose specific hardware which is not available outside the Louvre and requires that Bluetooth enabled hardware be installed in the gallery [18].

Klopfer et al. proposed a "location aware field guide" which operated in a manner similar to *Musing* but it was not adapted to use in a museum [19].

Lee et al. used an ultra-mobile PC, inertia tracker and camera for object recognition [20]. This system did not rely on external devices; instead, it relied on template matching. In this case, a translucent image of the next artwork is placed on the screen, guiding the user to the next artwork to be matched and used to locate the user within the museum space, attempting to estimate the user's location by the last artwork scanned. However, this approach does not provide for an accurate location estimate. Furthermore, this project relied on proprietary hardware supplied by the institution.

Another system that used specialized hardware to provide an augmented reality experience is described in [21]. The system overlays the picture of a physical image displayed on a custom hardware with pertinent information in real-time. The detection of the artwork is accomplished using ultrasound sensors and gyros for pose tracking. The information is then matched to the image using an edge-detection algorithm.

Explora-Museum-EXMU ([22]) is a tablet application that shares several features with Musing. It has a similar look and feel and similar client/server design approach[1]. Nevertheless, two key features distinguish the EXMU app from Musing. First, the app is currently available only on tablets. Second, and more important, the app requires special hardware in the form of blue tooth transmitters.

This might impose limitations on the flexibility of placement of artwork and rearrangement of the app upon changes in gallery / museum content.

[1] The application has been recently announced and there is no much information about it except for the information available in [22].

In addition to the previously discussed systems, there are a number of consumer-level museum applications that do not require proprietary hardware.

The Smithsonian Institution and Arcade Sunshine Media have developed *The Peacock Room Comes to America* app. The iOS application was built specifically to explore artist James McNeill Whistler's *Peacock Room* in the Smithsonian Freer Gallery [23]. The application allows for a virtual exploration of the space by presenting a scrolling image of the room with tapable artworks in the scene. When tapped, these artworks offer expanded textual and audio information. *Peacock Room* does not require the visitor to be physically located within the museum to view content, meaning that it does not actively drive visitors to the exhibit. The entirety of information (text, audio, and video) is locally stored on the user's device. As such, the application must be reprogrammed and downloaded again by the user, if information is edited or new information is created, which may result in the user missing updates and/or corrections/additions. *Musing* includes a setting referred to as the "Permanent Exhibition," which allows museums to create sampler exhibits to advertise new exhibitions. However, in addition to this option, *Musing's "*AR" optio enables augmented reality and real-time/on-location user interaction with the artworks on exhibition within the galleries.

The Museum of Modern Art (MoMA) in New York has developed the *MoMA* application, containing a large amount of information about the museum, including a calendar, ability to purchase tickets, and the ability to browse the MoMA's extensive collections, either by physically visiting the museum or browsing at home [24]. *MoMA*'s primary interface involves typing-in reference numbers (located next to artworks in the gallery) to allow visitors to listen to audio descriptions of artworks and view large photos. Much of the information is not locally stored on the device and is downloaded from an online database. Although there are reference numbers posted next to artworks in the physical gallery, the visitor is not required to visit the museum in order to consume the information. Additionally, content is not relayed contextually within the picture-plane which does not allow for direct connections to be made.

Reality Check, created by the McNay Museum of San Antonio, allows visitors to use their own device's (smartphone, tablet, etc.) camera (or that of a loaner device) to scan artworks in the physical gallery to initiate image recognition [25]. The application is designed to be game-like, allowing the visitor to recognize an artwork by first selecting a "clue." These clues are unique shapes of objects present in the artwork. Once the chosen shape is recognized in the artwork by the device's camera, the visitor is presented with supplemental textual, audio, and video information. *Reality Check* stores all of the information locally on the hardware, thus, a new build of the application is required as information is edited or created.

While the aforementioned systems show promise, they suffer from a variety of potentially problematic issues. Of the systems that require proprietary hardware, museums must use financial resources to purchase and maintain loaner devices. Systems that rely on external devices, such as Bluetooth emitters, increase workload of museum staff who must install them within the space. Most importantly, the majority of these systems require reprogramming when content is created and edited.

IV. Application Development and Deployment

Musing was developed by an interdisciplinary team that included researchers within computer science, communication design, and museology. The client application, built on iOS, was initially deployed from October 8th, 2013 through November 14th, 2013, in The University Galleries at Texas State University, for the exhibition, *Eric Zimmerman: West of the Hudson* (Figure 2) (additional example images, scanable by *Musing*, are available in [26]). During the 38-day run of the exhibit, 242 visitors downloaded *Musing*. In addition, 11 visitors checked-out iPod Touch devices provided by the galleries, indicating a high number of visitors used their personal devices. Gallery guest book logs showed that a minimum of 962 visitors attended the exhibit, denoting that about 25% of visitors had chosen to use *Musing*. This indicates a relatively strong initial acceptance rate of the concept. However, these figures do not account for repeat visitors, visitors who did not sign-in at the front desk, or visitors who shared devices.

Figure 2. *Head of State* by Eric Zimmerman, 2013. Example artwork from exhibit, *West of the Hudson*

The first deployment of the Musing client indicated promising results. However, data for the exhibit was manually input into the database by developers. In order to fully test a system that could be deployed in a functioning museum, the Web-based administrator panel would need to be tested as well.

A second deployment was designed to test the entire system, including the museum professional's ability to add, edit, and delete exhibit content with the Web-based *Musing Administrator Panel* (MAP). In addition, new artworks were chosen, which created unique challenges for the image recognition algorithm and were used in order to test its robustness.

In order to test for a greater variety of artwork media, the second trial utilized two concurrent exhibitions, which ran in two separate rooms of the gallery from March 17 through April 11, 2014. The first was an exhibition of photographs by artist, Lauren E. Simonutti titled, *The Devil's Alphabet*. The second was an exhibition of paintings by artist Richard Martinez titled, *¡PAINTINGSFORNOW!*, This exhibit was chosen explicitly because of the artworks' strong silhouettes, large areas of solid color, and limited visible surface detail.

Before exhibition installation and during content development, the museologist was able to input data into the database via the MAP for both exhibitions. This allowed the user to add, edit, and delete information, which included the uploading of reference photos, adding and rearranging POIs, populating content for the added POIs, and adding artists' biographical information. Additionally, this trial allowed the development team to discover any programmatic issues and resolve them during the data entry process.

Testing in the gallery indicated that the imagery in *The Devil's Alphabet* was satisfactorily recognizable by *Musing* (Figure 3). As these artworks were photographic prints behind glass, there were some adjustments needed for lighting within the exhibition space in order to minimize environmental reflections, which circumstantially interfered with image recognition.

Figure 3. *The Devil's Alphabet: A* by Lauren E. Simonutti, 2007. Example artwork from exhibit, *The Devil's Alphabet*

Musing's recognition rate of artworks in *¡PAINTINGSFORNOW!* was not satisfactory. As the paintings in this exhibit displayed strong silhouettes, but very little surface variation in tone or texture, it is theorized that the flat color and limited amount of detail in the artworks were the cause of the recognition failure (Figure 4). As an alternative, this exhibit was offered as a "Permanent Exhibit" within the *Musing* library so that the visitor could still access and view the information without utilizing image recognition. This points to a need to improve the image recognition capabilities of *Musing* for artworks of this kind.

During the second trial, additional 58 users downloaded *Musing*. This number is influenced by the fact that the second trial took place during the same exhibition schedule as well as the same exhibition venue. Visitor attendance logs establish the fact that because the venue is within an academic setting, many of the visitors are the same for each exhibition. As a result, it is thought that the majority of users may have already downloaded *Musing* for the prior usage.

Figure 4. *BEALDARC* by Richard Martinez, 2012. From exhibit, *¡PAINTINGSFORNOW!*

A. Pedagogical Design

Making associations is essential to deepening understanding and the pedagogical shifts that are occurring within museology reflect the changing needs of the museum visitor. In addition, art museums may have difficulties in identifying effective ways to provide the proper context for the art they exhibit, something that may result in a lack of connection to their visitors. As such, the use of *Musing* can result in an enriched aesthetic and educational experience for the visitor and provide a large context for exhibited artwork to encourage and deepen personal connections to the exhibition objects and expand the visitors' knowledge and understanding of the artwork, itself. These connections can be made by broadening the context for the novice viewer while adding to the experience of the initiated viewer. Further results can be a bridging of gallery programming within the daily life of the visitor via their in-gallery experience and connections. The use of *Musing* within an exhibition setting can provide an interpretive framework, which allows access to supplemental didactic information about the exhibitions while offering opportunity for interactivity.

At the heart of the concept of the ideal 21st century museum/gallery experience is what educator and innovator John Dewey referred to over a century ago when he spoke of the importance of interactivity to provide for an enriched learning environment [4]. Such interactivity, and the resulting enrichment, requires providing for diverse learning styles by including visual/print, visual/picture, auditory, and verbal/kinesthetic modalities, as well as a variety of user types [12]. These enriched learning environments are comprised of seeing, hearing, and interaction by moving beyond the traditional linear model of communication that provides didactic information via textual labels and gallery talks, to a non-linear model of communication through the provision of individual POIs, associated with each scanned artwork. Through the visitor's ability to access the POIs, which reflect a variety of types of didactic content contained within *Musing*, the application provides for an enriched environment in which the visitors can participate in creating a large context for the works exhibited. The provision of additional information about each work via POIs positions the visitor as a collaborator in the process of making meaning and serves to engage the visitor with the provided information which solidifies the content knowledge [4]. Meaning is made by the viewer in a variety of ways, which can begin by looking at art through several different filters. The individual POI provides an opportunity to show the viewer the works within an art historical, biographical, conceptual, or technical framework. As museums and galleries continue to seek ways in which the visitor's experience can be augmented, these POIs are an effective way to provide access for visitors to contextual information for the exhibited works, broadening the exhibitions' theses for the novice viewer as well as augmenting the meaning for the initiated viewer. This extends the application's ability to meet the needs of a variety of visitors who learn in different ways and access works on a multitude of levels, as well as John Falk's five types of user experiences [12]. As such, the broadening of the exhibited works' context via interviews, videos, Web sites, source material, art historical influences, and other art with shared conceptual frameworks allows for a personalization for the visitor through the implied narratives. This is thought to be the most effective way to expand the context for the work and deepen viewers' connections through the exercise and action of gathering the information [2]. The resulting associations within the gallery setting, moving into the viewers' world, are essential to deepening the understanding of subject matter—a result of the user transferring what he or she already knows and reflecting upon it [4].

For the novice viewer, whose frame of reference may be lacking in depth to fully make these associations, the POI format is ideal to expand reference points. As these associations and connections deepen, the experience begins to look familiar, something that can also make looking at art more comfortable. As Marjorie Schwarzer writes, "Today, when the meaning of art is more contested than ever,

[technologies] offer visitors the possibility of diverse interpretations" [27]. Schwarzer adds, "The branches of information available on these devices are close in spirit to the multiple ways in which we engage art" [27]. The ability to allow for different levels and a wide range of information, as well as a seemingly endless number of interpretive applications, reflects the diversity of the museum audience, itself [27].

Marjorie Schwarzer also notes, "As society is bombarded with rapidly changing multimedia messages, our ways of deciphering and understanding information have changed. We increasingly rely on a combination of sound, moving image, and text. Like it or not, new technologies outside of a museum's four walls alter the way that people process information inside the museum" [27]. *Musing's* effectiveness comes from the immediacy with which the user can access the POI content and making information available on demand allows for visitors to move freely within the space, not having to rely upon the preconceived schedule of their guide or any predetermined path.

Within the preferred postmodern approach to museology, the ability of the visitor to gain information and knowledge in an interactive capacity reflects several of the key tenets of the *New Museology*—value, meaning, and access—while allowing for greater meaning and relevance of the content in contemporary society [28]. An undesirable level of institutional authority can be implied or inferred through exhibitions that are authoritative in their approach to didactic display and interpretation, seen in limited interpretive labels and language wherein the curator's voice is solely represented. Without the constructed intellectual space needed to create meaning, the visitor may fail to foster an individual relationship with exhibition objects [28]. This, in turn, can determine whether the visitor's experience is enriched, aggregated, and circular in nature—comprised of many small connections formed between objects and the visitor's personal connections—or an isolated, linear-oriented experience—formed from objects considered in isolation via limited interactivity. As such, the visitor's relationship and connections to exhibition objects depends heavily on subjective and experiential aspects such as interactivity and consumption of information with which they make their own meaning [29]. We can see the ways in which visitors' relationships to objects are defined by how active/passive they are allowed to be; the more restrained the institutional authority associated with the experience is, the closer the relationship may be that the visitor can develop with the object [28][29].

The effects of this enriched experience build on each other. Providing a large narrative context for the exhibition objects allows the visitor to make greater connections with the individual works of art within an exhibition and make connections between the works contained within the exhibition and a large relationship between exhibitions offered through *Musing*. In this way, the artworks themselves become an interpretive tool, which allows for a familiar relationship on the part of the visitor and a greater connection to them. This focus on communication of content and provision of context for the object is what Stephen Weil refers to as "The Poetics and Politics of Representation" [28]. In so doing, the visitor looks *at* the featured works and sees, understands, and connects *through* them.

B. Client User Interface Design

Musing was designed to employ a client-server architecture that allows museum administrators to upload, remove, and alter content, post-deployment. This is accomplished through an administrative Web interface (MAP) which feeds the shared database. The application retrieves this content as requested by the user. This approach allows the material provided to the user to be as current as possible. Hence, the application is flexible and not limited to "on board" data, allowing any museum to closely serve the needs of its visitors. The application relies on an open source library called *OpenCV* for the processing and recognition of images which have been captured by the user.

The User Interface was designed in such a way as to adhere to the Apple Human Interface Guidelines for a tab-bar navigation style application: Consisting of the Exhibitions Screen, Scan Artwork Screen, Artwork View Screen, and Favorites Screen.

C. The Exhibitions Screen and the Artwork View Screen

The Exhibitions Screen, depicted in Figure 5a, consists of a list-view of exhibits that a visitor can visit, organized by "Permanent Exhibits" and "Augmented Reality Exhibits". Permanent Exhibits are previews of the experience that visitors can expect when using the application in-gallery. These exhibits contain artworks that can be viewed outside of the gallery setting (e.g., residence, dorm, etc.). This type of exhibit is included to advertise the application's features, to familiarize the user with the way that the application works, and encourage users to attend a live exhibition. The AR exhibition section includes exhibits that must be attended in person to view the didactic information for the artworks. This view provides information such as the name of the exhibit, in which museum the exhibit is located (provided more than one organization uses *Musing*), and a representative image to advertise the exhibition. Figure 5b shows a portion of the "Art View" screen: a captured and identified image along with the overlaid POIs.

POIs—tapable buttons that represent the types of content available to the user—are able to provide the user with a variety of didactic information. The individual POIs are as follows: 1) Factoids: Small pieces of text that can be attached to a feature in an artwork (Figure 6a); 2) Web site: Links provide information about the artist, or historically pertinent information (Figure 6b); 3) Video: Takes the user to an established internet video site such as YouTube and Vimeo or a locally hosted video within the application (Figure 7a).

These individual POIs can be tied directly to the aforementioned learning types (visual, auditory, and kinesthetic) written about by Pashleret et al. [11]. Through the diversity of information dissemination methods such as video, web based content, and text, as well as image based content, the visual learner's needs are met, while the auditory learner is stimulated as well by video and audio files and the kinesthetic learner enjoys the interactivity with the technology, itself. Through the exploration of the elements that comprise the experience provided by Musing, each of the learning types can be stimulated in ways that allow for their access to the content.

Figure 5. (a-top) Exhibitions Screen, including exhibition selection, and primary navigation; (b-bottom) A captured and identified image along with the overlaid POIs.

Figure 6. (a-top) Factoid POI; (b-bottom) External Web site

D. The Favorites Screen

Many museum visitors wish to retain information in order to consume or refer to at a later date. *Musing* allows the visitor, to favorite any of the artworks they scanned while visiting the museum. These favorites are saved in the Favorites Screen in a list view for later retrieval (Figure 7b).

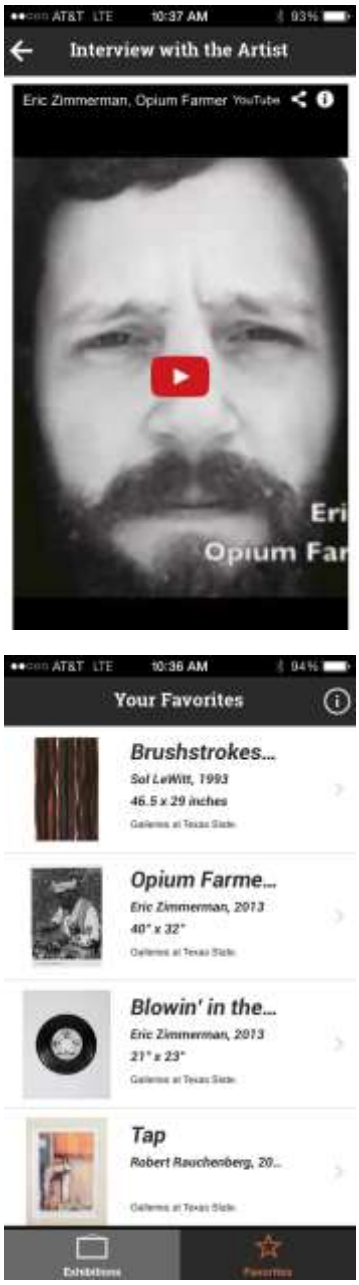


Figure 7. (a-top) YouTube video, created and uploaded by the musing professional; (b-bottom) Favourites Screen with list-view of saved artworks

E. Server User Interface Design

In order for *Musing* to be used in a wide variety of museums and galleries, the MAP Web site was created to provide museologists with the ability to easily create, retrieve, update and delete content in the system. As all consumable content for the client application is provided from a database, without MAP, the *Musing* system would require expensive upkeep by software developers.

MAP includes four pages for data entry: Exhibits, Artworks, Edit POIs, and Artists.

The Exhibits page allows the user to create/add exhibits, edit, and delete existing exhibits (Figure 8). From this page, the user is able to select existing exhibits for editing as well as create new ones.

Figure 8. Exhibits page, showing existing exhibits

When a new exhibition is created, MAP initiates the Edit Exhibition page (Figure 9). This interface allows the user to browse their local machine for an exhibition image (automatically resized by the system), choose a beginning and end date for the exhibit, enter the museum or gallery name, and set the exhibit type to Permanent or Augmented Reality. This information is displayed in the client on *Musing*'s Home Screen (Figure 5a).

Figure 9. Adding a new exhibit (detail)

After a new exhibit is created, the user is taken to the Artworks page. This page allows the user to add new artwork images to the exhibition, delete artworks, or edit artworks within the exhibition (Figure 10).

When adding a new artwork to an exhibition, the artwork editing page allows the user to upload and crop a reference photo of the artwork (used for image recognition by the client application) and enter information about the artwork. This information includes the artwork's title, dimensions, materials, year created, and artist (maintained separately by the Artist page). The entirety of this information is displayed in the *Musing* client after image recognition has taken place (Figures 5b and 6a).

From the selected artwork's page, the user is able to edit the POIs (Figure 11). The user has the ability to add new POIs, placing them by clicking and dragging. Additionally, the user is able to assign content to each POI,

and assign a category: History, Technique, Information, Web, or Video. The POIs are assigned (x, y) coordinates and appear in the *Musing* client in the same locations on the artworks (Figures 5b and 6a).

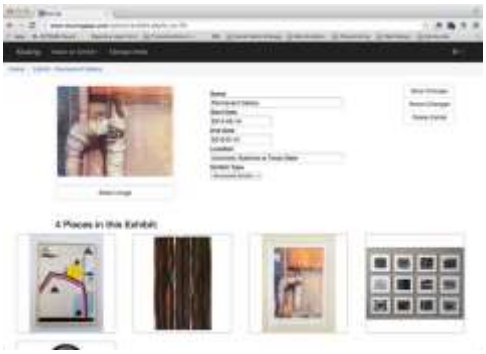


Figure 10. Managing artworks within an exhibit

Artist information is kept separate from the exhibits and the individual artworks to avoid duplication of data entry. The Artist section of MAP allows the user to add new artists or edit existing ones (Figure 12).

Figure 11. Editing POIs on the artwork (detail)

The Edit artist page allows the user to upload and crop a photo of the artist, input names, birth/death dates, and links to artist biographies, as well as bibliographic references. This biographic information is displayed in the *Musing* client at the bottom of the View Artworks Screen (Figure 5b).

Figure 12. Editing Artist information (detail)

After selecting an exhibit, the authenticated user is presented with a thumbnail for all of the artworks currently associated with that exhibit. In addition, the user is given the option of adding a new artwork to the exhibit. When a new work is added, the user selects an image of the art from local storage on their machine. The image is expected to be cropped such that only the artwork itself and its frame are shown. This greatly improves the recognition performance of *Musing* and creates a better experience for users of the application.

When an image has been selected for a new artwork, the user is directed to a page where information regarding the particular artwork can be entered or edited. This same screen is reached when an existing work of art is selected from the exhibit listing. The user is able to enter the artwork's title, size, year of creation, medium, and the artist's name. Artists are stored and catalogued in the database and information such as year of birth, year of death if applicable, and a link to a biography, can be entered and stored as a unique entry to the artwork in the database to avoid duplication of entries.

Next, the administrative support utility enables the administrator to define and edit POIs for an artwork. This is done using a graphical interface designed with JQuery. The user selects a position on a displayed image of the artwork, chooses the media type that the POIs references—along with its associated icon—and the text or URL as appropriate. In addition, users can alter the position of existing POIs by dragging and dropping them. The user can add and modify exhibits, as well as artists in a manner similar to that described for artworks.

F. Hardware/Software Architecture

The *Musing* server, or MAP, UI is constructed with HTML and CSS, reading from and writing to a MySQL database hosted on a Linux Web server. Currently, the *Musing* client runs on iOS based hardware, such as iPhone, iPod Touch, and iPad. An Android version is under design.

1) Back-end Processing

The back-end (server) application provides two main functionalities. First, it supplies information in the form of reference images and relevant didactics to the user, enabling its operation inside the gallery or with a permanent exhibition. Second, the back-end is designed to provide an administrator (e.g., a museum staff member) with the capability to edit the contents of an exhibition within the system. The server, which is shared by the application and the administrative support back-end utility, is used by the gallery administrators to load content into *Musing*.

The back-end, administered by MAP, is written in PHP and uses standard web-technologies (including HTML, CSS, JavaScript, AJAX, jQuery, and several Open-Source JavaScript libraries) to deliver a user-centric experience. It is designed to allow users unfamiliar with database systems to create, read, update, and delete entries for exhibits from a database stored within the web application's framework. The entries include artworks contained within a chosen exhibit, the associated artists, and curated POIs.

The primary vehicle for data entry into MAP is via Web forms depicted in Figures 8-12. These forms, when submitted, write data into the appropriate fields in the database. The database for the entire system is comprised of eight tables. One table is responsible for user authentication,

along with another which records failed login attempts. Two tables are responsible for tracking permissions of the exhibits and the artists. These tables separate exhibits and artists by user, so that administrators may only view their own information. The remainder of the tables are responsible for holding artworks, exhibits (Figures 9-10), POI placement/content (Figure 11) and data for artists (Figure 12).

When a new image is uploaded via a Web form, either for an artist headshot or an artwork reference image, the image is saved into a folder on the server and a pointer is saved to the appropriate database table for later retrieval.

JQuery and JavaScript are used to facilitate the placement of POIs (Figure 11), by allowing the museologist to drag and drop POIs wherever they wish in the picture plane. The (x, y) coordinates of the POIs are saved to the database in the POI table, along with the Artwork's ID, icon type, media type (e.g., text, video, audio) and URL for that content.

Musing was developed with the intention of packaging within the application as little data as possible. When the user activates *Musing*, it requests an XML document containing a list of available exhibits from the back-end data server. The application parses the XML document and extracts the information into an Exhibit object within the application. Along with the XML document, which contains the names of the exhibits, locations, and id values which the application can use to retrieve data about specific exhibits, the application retrieves a "banner image" for each exhibit, which is displayed in a list for the user to browse.

When the user selects an exhibit from the list on *Musing*'s Home screen (Figure 5a), the application passes its id value to a PHP script hosted on the data server. This process is referred to as 'synching'. During synching, the server compiles the pertinent information and returns information in the form of XML file and a set of JPEG images of the gallery artworks to the app. The XML document contains information about each artwork, along with the set of POIs related to the information. The user can tap on POIs to display additional information about the artwork or artist. The images retrieved along with this document are used both for displaying POIs on the Artwork View screen and as references by the image recognition.

As in the case of the exhibit list, the XML document provided by the data server when the application is synched to a particular exhibit is parsed. The extracted information is used to populate painting and POIs within the application for each painting and POIs listed in the database. The images are also incorporated into these objects. Testing has shown that this process of synchronization typically takes approximately 20 seconds, during which time the user is shown a modal progress graphic.

2) Front-end Processing

Musing supports two types of exhibits— permanent and AR. The synching process is the same for both. If the database indicates that an exhibit is permanent, the user is shown a list of artworks available in an exhibit and each may be selected by tapping. This displays the artwork's image with the proper set of overlaid POIs. The second type of exhibit is the AR variety. In this case, the user is given an image detection view rather than a list, which displays a real-time feed from the devices camera over which is laid a graphic of an empty painting frame, along with a button which the user can use to capture a photograph.

During image detection, the users are instructed to position themselves so that a *Musing* enabled artwork fully fills the frame displayed (this is not mandatory, yet it can improve the recognition rate) on the device's screen and to take a picture of the artwork. When this is done and an image is captured, the application compares the captured image to each reference image currently synchronized for the exhibit. If a match can be made, the application proceeds to the Artwork View screen, exactly as it does when the user selects an image in a permanent exhibit. Otherwise, an error message is displayed in a modal dialog. To save in storage space, the captured image is discarded after being matched or rejected.

From the Artwork View screen, the user has the option of capturing the artwork and its information by making the artwork one of their "Favorites." This is the only condition under which *Musing* locally stores the artwork and its information. This is done by passing the image, POIs data, and artist information to a Favorites Database object that incorporates those values into an array of artwork objects. The data is then written into *Musing's* internal database. The information stored in the favorites array is accessible by the user regardless of whether or not the device is connected to the internet.

3) Image Processing and Recognition

Musing relies on the Oriented FAST and Rotated BRIEF (ORB) image detection algorithm [30]. The ORB procedure combines the "FAST" key-point detection and "BRIEF" determination of descriptors. Key-points are clusters of pixels within an image which are unusual enough to stand out and to help distinguish a particular image from other images. After identifying a set of key-points within an image, a set of descriptors is calculated for each key-point using BRIEF [30]. This functionality is provided by the *OpenCV* open source computer vision library which is available for use in iOS and Android devices.

Key-point detectors frequently rely on finding "corners" and "edges" within images since image boundaries often create distinguishable pairings of shade and color [30]. ORB is translation invariant. Additional operations are performed to compensate for rotation and scaling [30].

In the training stage, BRIEF employs binary comparisons between pixels in a smoothed image [30]. This algorithm takes a relatively large set of key-points—often as many as 500—and builds a classification tree for the set. The tree serves as an image "signature" used to measure similarities between images. Alternatively, under the

approach used in this research, one can employ the results of the BRIEF stage using the *k* nearest neighbors (kNN) and one-to-one and onto mapping (bijection) test approach.

Following the synching process, users can point their device camera at an artwork in the gallery and capture its image. This image is processed using ORB and then compared to each of the reference images which were downloaded at sync time. Each reference image is processed to determine its key-points / descriptors at the time of comparison and this information is recalculated for each comparison. *Musing* employs the kNN and bijection approach to the key-points. Each key-point in a captured image is compared to each other in the reference image. A small set of matching key-points in the reference image is found for each key-point in the captured image. The goal is to find a maximal, high reliability, bijection between a subset of the key-points in a reference image and a subset of the key-points in the captured image. Hence, if any key-point in the reference image matches more than one key-point in the captured image with equal reliability, then *Musing* dismisses that match. The literature has suggested 0.65 as a reliability threshold and as the best threshold ratio for selecting one match as superior to the other [31]. Musing image recognition procedure uses this (0.65) threshold. The kNN is done twice, creating a set of directional matches that compares the reference image to the photograph taken and vice-versa. Then both sets are compared, dismissing any match that is not bidirectional. If a significant number of bidirectional matches is identified, the images are considered a match. *Musing* currently uses a threshold of 4 bidirectional matches as the minimum subset size.

When *Musing* has determined that a captured image matches a reference image, the reference image is displayed on screen along with an overlay of POIs.

The following is a description of the applied image recognition algorithm, starting with the captured image and the first reference image.

Step One: Captured Image Key-point Calculation - Find the key-points for the captured image using the FAST method [30]. This method checks a ring around each pixel and compares their intensities. It returns the point as a key-point if the gray level of a number of pixels within the ring is sufficiently higher or lower than the nucleus pixel itself.

Step Two: Captured Image Descriptor Calculation - BRIEF is used to take a patch of pixels surrounding a key-point and uses binary intensity thresholds to create a 256-bit binary vector describing the area around the key-point [30].

Steps Three & Four: Reference Key-points and Descriptors - Steps one and two are repeated for the reference image.

Step Five-A: Descriptor Matching (Captured to Reference) - A kNN matching of the Hamming Distances of each descriptor in the captured image to its K nearest neighbors in the reference image is performed. The two best matches for each key-point are retained.

Step Five-B: Descriptor Matching (Reference to Captured) - Step Five-A is applied with the roles of the captured and reference image reversed.

Step Six-A: Ratio-Test (Captured to Reference) - This step discards every match identified for the captured image where the best match and second-best match have similar Hamming distances. This produces a one-to-one match.

Step Six-B: Ratio-Test (Reference to Captured) - Weeding, using the same criteria as in step Six-A is performed on any match from the set of matches identified in the reference image.

Step Seven: Symmetry Cross-Check Test - The Symmetry cross-check test returns only the pairs of matches that are found from the captured image to the reference image and from reference image to the captured image. This process enables keeping only the strongest symmetric correspondences and maintaining a bijection.

Figure 13 illustrates the process performed in steps 5 to 7.

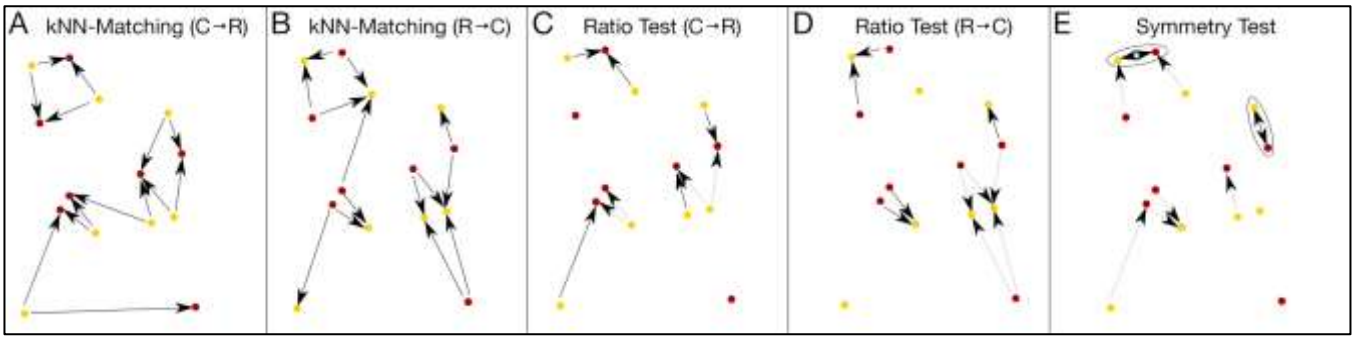


Figure 13. (A) and (B) kNN matching ($k = 2$); (C) and (D) Descriptor matching - the process discards matches with similar quality (Hamming distance) and retains the best match for distinctive matches; (E) Symmetry cross checking – only bidirectional matches are retained.

Step Eight: Output if Found - If four or more matches remain after the weeding performed by the ratio tests and symmetry test, the procedure retains the identity of the reference image and returns to step three for the next reference image (if such an image is available). The procedure keeps track of the identity of the image that produced the largest number of matches and outputs its id. If all reference images have been tested and no match has been found, then a message "Image Not Found" along with instructions to the user on ways for improving the possibility of match are displayed.

G. Design of Testing Instruments

Although this project seeks to augment the viewers' experience, while traditional didactics and tours remain in place, it is not simply an "add-on" to the material that the galleries already provide. It is a model for directional movement in museum practices, so to assure its success, proper analysis must be done. Utilizing Scott Sayre's model of evaluation—Pre-production Surveys, Formative Testing, Summative Evaluation, Audience Focus Groups, and Computer-collected data – we can get a thorough evaluation before, during, and after production that provides a myriad of benefits relative to the assurance of effectiveness [2].

Testing instruments consisted of quantitative benchmark testing and a qualitative user perception exit questionnaire. As a part of the quantitative testing, each reference and captured image has been processed to generate 500 identifying key-points in each of 60 total images. The 60 images consist of: ten reference images ($R_1 - R_{10}$) and ten images that served as captured images($P_1 - P_{10}$). Each of the captured images was captured four additional times for a total of five capturing per image. The first one used maximum alignment to the reference images the rest of the four were taken with increasing rotation translation and scaling (due to different distance). The maximal rotation was 40 degrees.

The procedure described above was applied to the ten reference images and fifty captured images. A threshold of 0.3% over the percent of matching key-points, which was empirically identified as the most suitable threshold was used by the program and applied to the matching results.

For the qualitative testing we have used a 23-question exit questionnaire designed to capture feedback from in-gallery users. The questions were written to determine the user's acceptance of the application, their perceptions of application performance, enjoyment of the application, as well as pedagogical concerns.

V. Deployment Results

A. Technical Results (Internal Testing)

Figure 14 shows a heat-map of the results of this experiment in the form of a confusion matrix. The figure shows a recognition rate of 96.4% with 0% error of type-1 (false positive) and 3.3% error of type-2 (false negative) obtained with $P_{(1,4)}$ and $P_{(1,5)}$. We have found however, that with rotation of more than 45 degrees there were numerous false negatives; but, still 0% of false positive error.

The testing has shown that *Musing* recognizes images with near perfect reliability under ideal conditions, that is, when a user is directly in front of the artwork, has positioned the artwork correctly within the image capture frame, and is not holding the device at an angle. Nevertheless, excessive rotation of the camera while capturing an image diminishes reliability. Our testing indicates that *Musing* recognizes images at a 45 degree rotation with 90% reliability and a 90 degree rotation with 84% reliability. The application performance degrades when the user stands off of the center line when photographing a piece of art, producing a skewed image. A slight deviation from the center (approximately 15 degrees) produced no noticeable change in testing but at greater values (approximately 45 degrees) the system produces 40% true positives and 60% false negatives. As far as can be determined, in the field-deployment testing, the system did not generate false positive results. Furthermore, the user surveys have not indicated that the application has produced a false positive error in use. Additionally, if the user stands too far from the artwork to properly fill the capture frame the reliability has suffered as well, with the reliability rate dropping to 48% at approximately twice the recommended distance.

Testing indicated that the image recognition algorithm failed when artworks were behind glass, causing heavy reflections, as well as those artworks with little tonal variation or surface detail. Artworks behind glass can often create reflections of the user as they are standing in front of the artwork. These reflections interfere with the image recognition by creating an image that falls outside the tolerance range of the algorithm. Artworks that exhibited little tonal variation (i.e., large patches of solid color) or little surface detail also created challenges for the image recognition algorithm, as there were not enough unique identifier points for the algorithm to affect recognition.

Testing performed to evaluate the processing time revealed that with 10 reference images, the application was able to compare and either display or reject an image in approximately 3.3 seconds on a stock iPod Touch-5. Again, user surveys indicate that this was sufficient to produce a positive experience for most users.

Finally, User surveys conducted during the trials indicate that the application's reliability was sufficient to produce a positive experience for most users.

B. Exit Questionnaire Results with Live Users

Of the pertinent questions, 83.6% responded that *Musing* was able to recognize the artwork "every time" or "most of the time." 77.5% considered *Musing* to be quick and responsive. 87.7% considered *Musing* enjoyable to use and 93.8% wishing to see *Musing* in a future exhibit.

VI. Results Evaluation

The deployment results show high recognition accuracy and relatively short synching/recognition delay time, therefore the functionality of the entire system has been

verified. The application has passed the Apple approval process and is available for download [5].

Formal user feedback obtained via questionnaire was consistent with our evaluation of the system and with informal feedback. The visitors that have responded to the survey have found the application as informative and usable. Their perception of precision and timing was favorable and overall they have commended the system and expressed interest in its further use. Informal feedback from users, including several staff members of other galleries, was overwhelmingly positive.

VII. CONCLUSIONS AND FUTURE RESEARCH

We have designed, implemented, and deployed a usable mobile application that facilitates an enriched museum visitor experience via AR using interactive didactics. Per our assessment, the application has achieved its stated goals and has shown that the research hypothesis is valid.

Although tried and true wall labels, pamphlets, and gallery talks are sufficient for conveying information and serve to extend interpretive opportunities [32], they carry with them constraints that do not adapt in the ways that mobile media can. Mobile media technology provides the ability to allow for different levels and a wide range of information, as well as a seemingly endless number of interpretive applications and these interpretive strategies can reflect the diversity of the museum audience, itself [27]. These diverse and changing multimedia messages are reflected in the ways that *Musing* can be used.

Ultimately, the knowledge and deepened understanding that *Musing* can facilitate is filtered through the learning and innovation skills of the 21st Century – that of creativity and innovation, communication and collaboration, and cross-disciplinary thinking [32].

The field testing via the exhibition shows that *Musing* can be used on non-proprietary smartphone hardware and provide visitors with didactic information, without the need for external tokens and reprogramming for information changes. This enables reduced reliance on loaner hardware. In addition, the implementation of MAP allowed for the museologist to curate an exhibition within a simple to use Web application, without the need for software development abilities. This ability allows musing to be deployed in external museums and galleries as a complete, turn-key solution.

A. Future Research

Future enhancements to the *Musing* smartphone application (client) will include abilities for users to share images and didactics via social media such as *Facebook* and Twitter, as well as the ability to comment on artworks within the application so users can "join in the conversation." Additionally, there are plans to complete a port of the current iOS-based implementation to the Android environment.

It was determined that number of user downloads did not provide sufficient information about the way that users were interacting with the client. The addition of data analytics within the client, including collecting the number of times a user accessed the client, the number of times that an artworks scanned, the number of POIs accessed, and additional information could provide insight into the relative success of the client.

Future research with regard to MAP includes a redesigned GUI and improved user experience, as well as user testing with multiple users in order to provide the research team with a plan for feature enhancements.

Other plans for future activities include expanding the image processing capabilities by further improving recognition accuracy, resilience, and time performance. We plan to investigate the integration of algorithms for recognition of 3-D objects using the smartphone/tablet camera.

Lastly, in the Fall of 2014 and Spring of 2015 the research team has tested the *Musing* system (client and server) in external galleries, not directly attached to this project. The Bluestar Gallery in San Antonio, Texas and the Wittliff Collection Gallery in the Texas State University has tested an exhibit with *Musing*. Both galleries provided important positive feedback concerning the experience of visitors that used the app and the ease of use for gallery stuff. Additionally, the feedback provided helped improving some of *Musing* features. We plan to continue testing *Musing* in external galleries. It is hoped that more information can be gleaned from implementing Musing from these exhibits and in a large variety of gallery spaces.

ACKNOWLEDGEMENT

The research team would like to thank Texas State University's Research Enhancement Grant program for providing the initial funding for this research. In addition, continued project funding was provided by the office of the Vice President of Research (Dr. Michael Blanda), Dr. Timmothy Mottet, Dean of the College of Fine Arts and Communication, Dr. Stephen Seidman, Dean of the College of Science and Engineering, Mr. Michael Niblett, Director of the School of Art and Design, and Dr. Hongchi Shi, Chair of the Department of Computer Science of Texas State University.

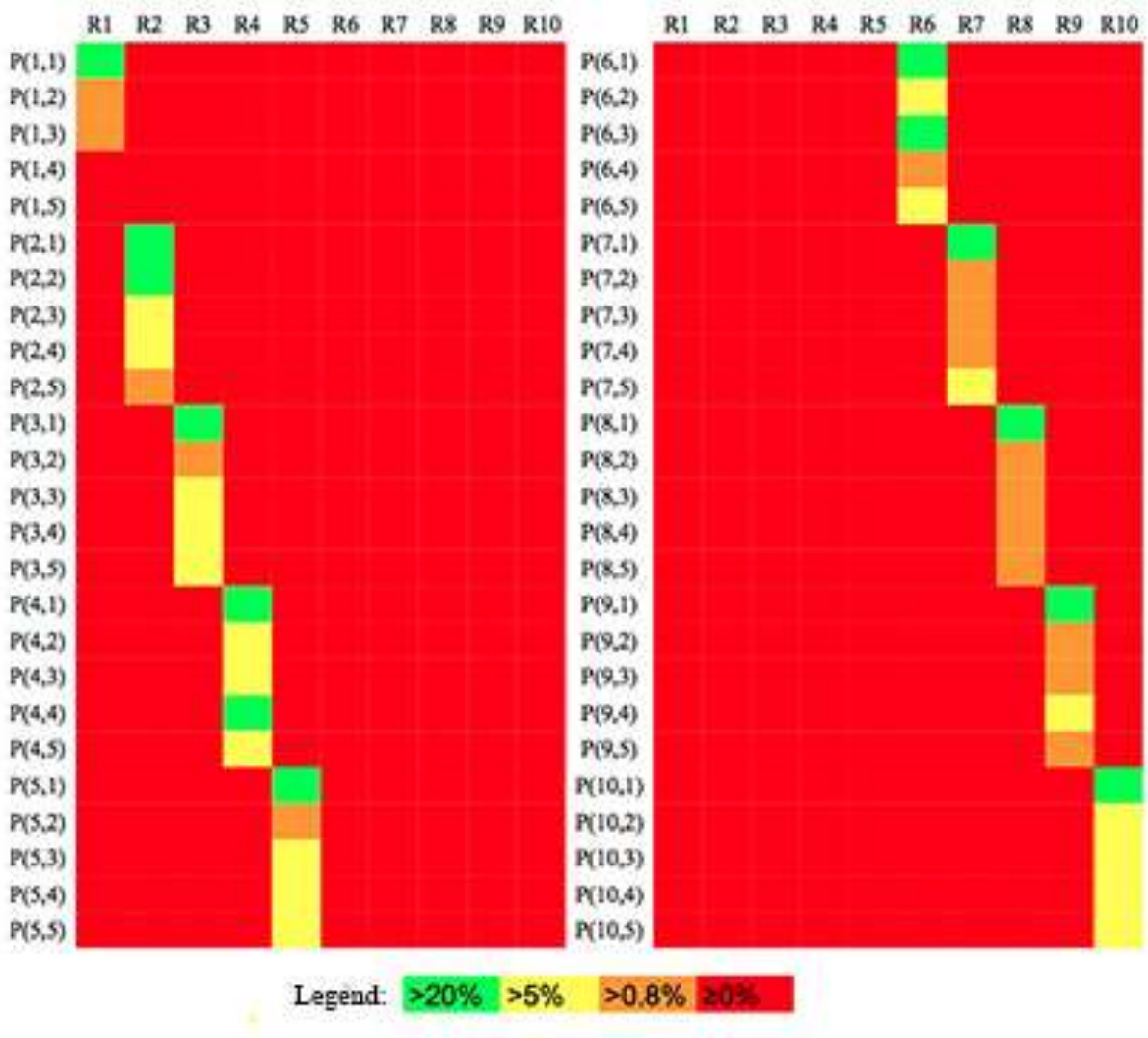


Figure 14. A heat-map of the results of the image matching experiment.

REFERENCES

[1] G. Atkinson, K. Whiteside, D. E. Tamir, G. Lawrence, and M. M. Stump, "Musing: Interactive Didactics for Art Museums and Galleries via Image Processing and Augmented Reality - Providing Contextual Information for Artworks via Consumer-Level Mobile Devices," proceedings of the 6th *International Conference on Creative Content Technology, Content 2014*, Venice, Italy, May, 2014.

[2] S. Sayre, "Assuring the Successful Integration of Multimedia Technology in an Art Museum Environment," in S. Thomas and A. Mintz (Eds.), the Virtual and the Real: Media in the Museum, Washington, D.C., 1998, pp. 1-10.

[3] K. Morrissey and D. Worts, "A Place for the Muses? Negotiating the Role of Technology in Museums," in S. Thomas and A. Mintz (Eds.), the Virtual and the Real: Media in the Museum, Washington, D.C., 1998, pp. 147-171.

[4] T. C. Clapper, "The Enriched Environment: Making Multiple Connections" in the Academic Leadership Journal, 8(4), 2010, pp. 1-2.

[5] *Musina*, a photo recognition application that allows users to scan artwork at participating museums and art galleries to learn more about the work, Apple Store, https://itunes.apple.com/us/app/musing/id694382407?ls=1&mt=8, [retrieved August 2014].

[6] G. Lawrence and M. Stump, "Connecting Physical and Digital Worlds. A Case Study of Quick Response Codes and Social Media in a Gallery Setting," The International Journal of Design in Society, 6(3), 2013, pp. 79-95.

[7] G. Atkinson, K. Whiteside, G. Lawrence, M. M. Stump, and D. E. Tamir, "Musing: Enhancing Educational Experience through an Augmented Reality/Virtual Museum Application," *in proceedings of the, 8th International Technology, Education and Development Conference*, Seville, Spain, March, 2014.

[8] K. Whiteside, G. Atkinson, M. M. Stump, G. Lawrence, D. E. Tamir, "Musing: Adaptable Mobile Augmented Reality Application for Museums and Art Galleries," in *proceedings of the Electronic Visualization and the Arts, EVA 2014*, London, UK, July, 2014

[9] L. Tallon, "Introduction: mobile digital and person" In L. Tallon and K. Walker (Eds.), *Digital Technologies and the Museum Experience: Handheld Guides and Other Media*, Lanham, MD; AltaMira Press, 2008, pp. xiii-xxv.

[10] S.E. Weil, "Making Museums Matter" Smithsonian Institution, Washington, D.C., USA, 2002.

[11] H. Pashler, M. McDaniel, D. Rohrer, and R. Bjork, Learning styles: Concepts and evidence. *Psychological Science in the Public Interest 9*, 2009, pp. 105–119.

[12] J. Falk, *Identity and the Museum Experience*. Walnut Creek, California: Left Coast Press, 2009.

[13] O. Bimber, and E. Bruns, "PhoneGuide: Adaptive Image Classification for Mobile Museum Guidance," IEEE International Symposium on Ubiquitous Virtual Reality, Jeju, South Korea, 2011, pp.1-4.

[14] M. Hatala, L. Kalantari, R. Wakkary, and K. Newby, "Ontology And Rule Based Retrieval Of Sound Objects In Augmented Audio Reality System For Museum Visitors." *ACM symposium on Applied computing*, New York, NY, 2004, pp. 1045-1050.

[15] C. Jing, G. Junwei, and W. Yongtian, "Mobile Augmented REality System For Personal Museum Tour Guide Applications". IET Wireless and Mobile Computing, Shanghai, China, 2011, pp. 262 – 265.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an Efficient Alternative to SIFT or SURF" IEEE International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2564-2571.

[17] M. Blöckner, S. Danti, J. Forrai, G. Broll, and A. De Luca, "Please Touch the Exhibits!: Using NFC-based Interaction for Exploring a Museum." *International Conference on Human-Computer Interaction with Mobile Devices and Services*, New York, 2011, Article 71.

[18] T. Miyashitat, et al., "An Augmented REality Museum Guide, in IEEE International Symposium on Mixed and Augmented Reality, Cambridge, 2008, pp. 103 – 106.

[19] E. Klopfer and K. Squire, "Environmental Detectives—The Development of an Augmented Reality Platform for Environmental Simulations," in Educational Technology Research and Development, 56(2), 2008, pp.203-228.

[20] D. Lee, and J. Park, "Augmented Reality based Museum Guidance System for Selective Viewings," IEEE Workshop on Digital Media and its Application in Museum & Heritage, 2007, Chongign, China, pp. 379-382.

[21] J. Oh et al., "Efficient Mobile Museum Guidance System Using Augmented Reality," IEEE International Symposium on Consumer Electronics, Vilamoura, Portugal, 2008, pp.1,4, 14-16.

[22] Transform your museum visit into an incredible exploration http://www.explora-museum.com/ [retrieved, March, 2015]

[23] Peacock Room Comes to America iPhone and iPad application. https://itunes.apple.com/us/app/peacock-room-comes-to-america/id671150763?mt=8 [retrieved, March, 2015].

[24] MoMA iPhone application. https://itunes.apple.com/us/app/moma/id383990455?mt=8 [retrieved, March, 2015].

[25] Reality Check iPhone application. https://itunes.apple.com/us/app/themcnay-reality-check/id615135643?mt=8 [retrieved, March, 2015].

[26] *Eric Zimmerman: West of the Hudson*, example images, scanable by *Musing*, http://www.musingapp.com/test_images/, [retrieved, March, 2014].

[27] M. Schwarzer, "Art & Gadgetry: The Future of the Museum Visit", Museum News. http://www.aamus.org/pubs/mn/MN_JA01_ArtGadgetry.cfm, [retrieved, March, 2015].

[28] D.C. Stam, "The informed muse: The implications of 'The New Museology' for museum practice," In G. Corsane (Ed.), *Heritage, Museums and Galleries*. London and New York: Routledge, 2005.

[29] A. McClellan, "Ideals and Mission," *The Art Museum from Boullee to Bilbao*, University of California Press, pp. 13-54, 2008.

[30] M. Calonder et al., "BRIEF: Computing a Local Binary Descriptor Very Fast," IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(7), 2012, pp. 1281-1298.

[31] E. Rosten, R. Porter and T. Drummond, "Faster and Better: A Machine Learning Approach to Corner Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1), 2010, pp. 105-119.

[32] Museums & Society 2034: Trends and Potential Futures, Center for the Future of Museums, American Association of Museums, Washington, DC http://resource.aaslh.org/view/museums-and-society-2034-trends-and-potential-futures-report/ [retrieved, March, 2015].

IQ Imbalance in Heterodyne Transceivers with zero-second-IF for Wide-Band mmW Links

Ainhoa Rezola*†, Juan Francisco Sevillano*†, Martin Leyh‡, Moises Lorenzo‡, Roc Berenguer*†, Aharon Vargas‡ and Igone Vélez*†

*Electronics and Communications Department,
Centro de Estudios e Investigaciones Técnicas (CEIT), 20018 San Sebastián, SPAIN
Email: {argarciandia,jfsevillano,rberenguer,ivelez}@ceit.es
†Electrical, Electronic and Control Engineering Department,
Technology Campus of University of Navarra (TECNUN), 20018 San Sebastián, SPAIN
‡Fraunhofer Institute for Integrated Circuits (IIS), D-91058 Erlangen, Germany
Email: {martin.leyh,moises.lorenzo,aharon.vargas}@iis.fraunhofer.de

***Abstract*—Millimeter wave links are an attractive solution for mobile network backhaul. In order to cope with the requirements of future networks, these millimeter wave links should achieve Gigabit data rates. These data rates can be achieved by using wide-band and high order modulations in E-Band. Heterodyne architectures are good candidates for integrated transceivers, but, the design of integrated transceivers at these frequencies is a challenging issue. An important source of degradation is I/Q imbalance, which can significantly reduce the performance of a communication system with zero-second-IF transceivers if it is not appropriately compensated. The article analyzes the source of this IQ imbalance and proposes the use of different digital processing techniques, including a linear adaptive equalizer scheme. The performance of the transceiver is analyzed at system level by means of simulations. The results presented in the article suggest that the use of those techniques is able to mitigate the impact of the IQ imbalance effects, in order to allow the use of a high-order modulation such as 64QAM.**

***Index Terms*—Mobile backhaul; millimeter-wave; transceivers; RF impairments.**

I. Introduction

The growing demand for ubiquitous broadband communication, e.g., fourth-generation (4G) wireless, has motivated the deployment of ultra high-speed communication systems. Particularly in backhauling networks, optical fiber is required to transport very high data rates. However, optical fiber exhibits important drawbacks, such as high costs, long deployment times, and low flexibility. Recently, point-to-point wireless communication systems have been proposed as an attractive alternative to optical fiber. In order to achieve data rates that are comparable to optical fiber, these communication systems demand very high bandwidth in order to transport enough data. Although the frequency spectrum is congested, the regulation of the E-band, which is around 80 GHz, facilitates the deployment of high-speed communication systems in which a huge amount of data can be transmitted. The European Telecommunications Standards Institute (ETSI) is carrying out a standardization process for this frequency band [1], [2], [3].

Commercial off-the-shelf communication systems operating in the E-band support data rates of up to 2.5 Gbit/s. However, new applications demand even higher data rates (around 10 Gbits/s), which necessitates both wide-band and high-order modulations to utilize the spectrum efficiently. High bandwidths and higher-order modulation waveforms pose special challenges for communication systems offering reliable links at very high data rates. The digital base-band must be able to encode/decode and modulate/demodulate a huge amount of information while employing high efficiency algorithms to guarantee reliable communications. In addition, analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) must work at very high sampling rates. Furthermore, higher order modulations are desirable to boost up the data rate, requiring higher carrier-to-interference (C/I) ratios at the receiver. This involves carefully analyzing the degradation effects introduced by the analog RF impairments and evaluating the corresponding compensation algorithms in the digital baseband processing [4].

In this article we focus on the IQ imbalance impairment, which is one of the performance-critical effects of interest in zero-second-IF transceiver architectures. This impairment, caused by mismatches in the amplitude and phase responses of the I and Q signal paths, entails a degradation in the Image Rejection Ratio (IRR), which is theoretically infinite and causes interfering images at mirror frequencies. IQ imbalance encountered in narrow-band systems can be regarded as non-frequency-selective (NFS), and it is mainly caused by the local oscillators used for quadrature modulation or demodulation. However, IQ imbalance in wide-band systems may also exhibit frequency-dependent or frequency-selective (FS) behavior due to mismatches between the analog filtering paths of the I and Q components caused by finite tolerances [5] [6].

The article analyzes the impact of both non-frequency-selective and frequency-selective imbalances in the transmitted and received signal. In addition, different options for mitigating the NFS and FS IQ imbalance are described. These techniques, based on digital signal processing, are evaluated in a 64-QAM transceiver operating with a signal bandwidth of 2GHz.

The remainder of this article is structured as follows. Section II provides a description of the system architecture.

Section III introduces the NFS IQ imbalance issue, identifying and modeling the source of this impairment. The mitigation of the IQ imbalance impairment by digital signal processing at the receiver is also described in this section. In Section IV, a simulation model for the transceiver system, including the proposed mechanisms for reducing NFS IQ imbalance, is presented and analyzed. In Section V, the FS IQ imbalance issue is addressed and in Section VI the simulation results regarding this impairment are presented. Finally, some conclusions are drawn in Section VII.

II. System Architecture

In order to address new applications for the future backhauling networks, a point-to-point microwave link in the E-Band using a 64-QAM modulation with a signal bandwidth of 2GHz is considered. Fig. 1 shows the proposed transceiver (TRx) architecture for a point-to-point microwave link in the E-Band. As shown, the transmitter (Tx) front-end consists of an IQ up-converting modulator that up-converts the baseband I and Q channels to an intermediate frequency (IF). After combining the I and Q channels, the IF signal is up-converted by means of the millimeter-wave (mmW) mixer. Finally, the wideband mmW power amplifier (PA) is used to amplify and transmit the mmW signal. The receiver (Rx) front-end consists of a wideband Low Noise Amplifier (LNA), which receives and amplifies the signal at the E-Band. After the LNA, a first mixer down-converts the mmW signal to the same IF as in the Tx. This way, the same PLL can be re-used for the Tx and the Rx. Finally, an IQ demodulator down-converts the IF signal to 0-Hz.

This architecture presents a good balance between different design aspects, and it enables the minimization of the sampling frequency of the digital-to-analog (DAC) and analog-to-digital (ADC) converters. Nowadays, we can find commercial DACs and ADCs able to provide sampling rates close to 3Gsps, which is enough for practical implementation of the zero-second-IF architecture.

Due to the high channel bandwidth (higher than 2 GHz), the architecture depicted in Fig. 1 presents a good balance between the DAC and the ADC requirements and complexity on the transceiver. The use of other architectures, such as non-zero-second-IF would require very high performance ADCs or DACs, with sampling rates well above 4 Gsps to achieve a practical implementation of base-band and image rejection filters in the analog front-end. Other IF architectures, for example [7], relax the performance of the ADCs and DACs by using several sub-bands, but it is at the expense of increasing the complexity of the baseband and IF sections of the front-end of the transceiver.

However, the use of a zero-second-IF architecture presents well-known issues that should be addressed in order not to degrade the performance of the transceiver. This architecture is subject to corruption due to IQ imbalances at both the transmitter quadrature modulator and at the receiver demodulator . The resulting system performance degradation can be significant, especially for high-order modulation schemes [8].

III. Non-Frequency-selective IQ Imbalance

A. Imbalance analysis

The goal of the IQ modulator in Fig. 1 is to perform a frequency translation of the signal. That is, if the base-band input signal to the IQ modulator is

$$\tilde{s}(t) = s_I(t) + j s_Q(t), \tag{1}$$

where $s_I(t)$ is the signal in the I-datapath and $s_Q(t)$ is the signal in the Q-datapath, a perfect IQ modulation mixes the base-band input signal with

$$l_{\mathrm{tx}}(t) = e^{j\omega_{\mathrm{tx}}t} = \cos(\omega_{\mathrm{tx}}t) + j\sin(\omega_{\mathrm{tx}}t) \tag{2}$$

producing an output signal

$$s(t) = \mathrm{Re}\left[\tilde{s}(t) l_{\mathrm{tx}}(t)\right] = s_I(t)\cos(\omega_{\mathrm{tx}}t) - s_Q(t)\sin(\omega_{\mathrm{tx}}t). \tag{3}$$

However, when implementing an IQ modulator with actual electronic circuits, the signals produced by the local oscillator (LO) will present some difference in their amplitudes and will not have a phase difference of $\pi/2$. The effect of this imbalance can be modeled as the mixing of the base-band input signal with

$$l_{\mathrm{tx}}(t) = \cos(\omega_{\mathrm{tx}}t) + j g_{\mathrm{tx}} \sin(\omega_{\mathrm{tx}}t + \phi_{\mathrm{tx}}) \tag{4}$$

to yield the output signal

$$s(t) = (s_I(t) - g_{\mathrm{tx}}\sin(\phi_{\mathrm{tx}}) s_Q(t))\cos(\omega_{\mathrm{tx}}t) \tag{5a}$$

$$- s_Q(t) g_{\mathrm{tx}} \cos(\phi_{\mathrm{tx}})\sin(\omega_{\mathrm{tx}}t). \tag{5b}$$

In IQ imbalance analyses, it is common to rewrite the signal produced at the transmitter LO with the imbalance from equation (4) in the form of [9]

$$l_{\mathrm{tx}}(t) = C_1 e^{j\omega_{\mathrm{tx}}t} + C_2 e^{-j\omega_{\mathrm{tx}}t}, \tag{6}$$

with

$$C_1 = \frac{1 + g_{\mathrm{tx}} e^{j\phi_{\mathrm{tx}}}}{2} \tag{7a}$$

$$C_2 = \frac{1 - g_{\mathrm{tx}} e^{-j\phi_{\mathrm{tx}}}}{2} \tag{7b}$$

and the transmitted signal is

$$s(t) = \mathrm{Re}\left[\tilde{s}(t)\left(C_1 e^{j\omega_{\mathrm{tx}}t} + C_2 e^{-j\omega_{\mathrm{tx}}t}\right)\right]. \tag{8}$$

In (8), the desired term is the one multiplied by $e^{j\omega_{\mathrm{tx}}t}$ and the term multiplied by $e^{-j\omega_{\mathrm{tx}}t}$ is considered an undesired image. Equation (8) can be rewritten as

$$s(t) = \frac{1}{2}\left(C_1 \tilde{s}(t) e^{j\omega_{\mathrm{tx}}t} + C_1^* \tilde{s}^*(t) e^{-j\omega_{\mathrm{tx}}t}\right) \tag{9a}$$

$$+ \frac{1}{2}\left(C_2^* \tilde{s}^*(t) e^{j\omega_{\mathrm{tx}}t} + C_2 \tilde{s}(t) e^{-j\omega_{\mathrm{tx}}t}\right), \tag{9b}$$

where $(\cdot)^*$ denotes the complex conjugate.

Let $X(\omega)$ denote the Fourier Transform of a signal $x(t)$, then from (9) we have

$$S(\omega) = \frac{1}{2}\Big(C_1 \tilde{S}(\omega - \omega_{\mathrm{tx}}) + C_1^* \tilde{S}^*(\omega + \omega_{\mathrm{tx}}) \tag{10a}$$

$$+ C_2^* \tilde{S}^*(\omega - \omega_{\mathrm{tx}}) + C_2 \tilde{S}(\omega + \omega_{\mathrm{tx}})\Big). \tag{10b}$$

Fig. 2 illustrates the spectrum of $s(t)$. The first line in (10) is the desired term and the second line is an image that is an

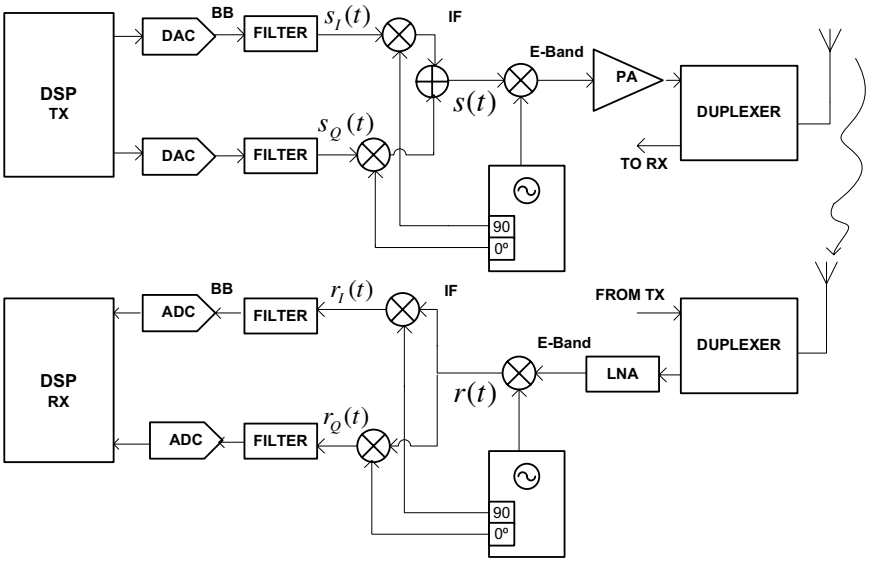


Fig. 1. Architecture of the transceiver.

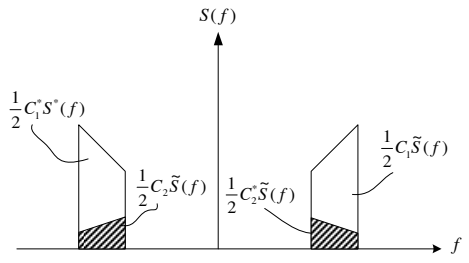


Fig. 2. Spectrum of $s(t)$ with imbalance.

alias of the desired signal. A measure of performance of the IQ modulator is the image rejection factor (IRR)

$$\mathrm{IRR}_{\mathrm{tx}} = \frac{|C_1|^2}{|C_2|^2} = \frac{1 + g_{\mathrm{tx}}^2 + 2g_{\mathrm{tx}}\cos(\phi_{\mathrm{tx}})}{1 + g_{\mathrm{tx}}^2 - 2g_{\mathrm{tx}}\cos(\phi_{\mathrm{tx}})}. \quad (11)$$

Similarly, the task of the Rx IQ demodulator in Fig. 1 is to mix the input signal with

$$l_{\mathrm{rx}}(t) = e^{-j\omega_{\mathrm{rx}}t} = \cos(\omega_{\mathrm{rx}}t) - j\sin(\omega_{\mathrm{rx}}t), \quad (12)$$

so that after low-pass filtering, the base-band equivalent of the signal in the frequency band of interest $\tilde{z}(t) = z_I(t) + jz_Q(t)$ is obtained. Note that $\tilde{z}(t)$ is the base-band equivalent that is referred to a carrier frequency ω_{rx}. The current implementation of the IQ demodulator will introduce similar imbalances to the actual implementation of the IQ modulator, which can be modeled as the mixing of $r(t)$ with

$$l_{\mathrm{rx}}(t) = \cos(\omega_{\mathrm{rx}}t) - jg_{\mathrm{rx}}\sin(\omega_{\mathrm{rx}}t + \phi_{\mathrm{rx}}). \quad (13)$$

The signal at the output of the IQ demodulator after low-pass filtering can be written as [9]

$$\tilde{r}(t) = r_I(t) + jr_Q(t) = K_1\tilde{z}(t) + K_2\tilde{z}^*(t) \quad (14)$$

with

$$K_1 = \frac{1 + g_{\mathrm{rx}}e^{-j\phi_{\mathrm{rx}}}}{2} \quad (15a)$$

$$K_2 = \frac{1 - g_{\mathrm{rx}}e^{j\phi_{\mathrm{rx}}}}{2}. \quad (15b)$$

In this case, the first term in the sum of (14) is the desired term and the second one is the image that aliases on the desired signal. The IRR for the IQ demodulator is defined as

$$\mathrm{IRR}_{\mathrm{rx}} = \frac{|K_1|^2}{|K_2|^2} = \frac{1 + g_{\mathrm{rx}}^2 + 2g_{\mathrm{rx}}\cos(\phi_{\mathrm{rx}})}{1 + g_{\mathrm{rx}}^2 - 2g_{\mathrm{rx}}\cos(\phi_{\mathrm{rx}})}. \quad (16)$$

Note that the above model for impairment affects to the whole information bearing signal in the same way. Although the model has been developed from the point of view of the local oscillators of the IQ modulator and demodulator, it can also be used to include the mean imbalances between the in-phase and quadrature datapaths.

In order to gain some insight, we assume enough linearity and proper filtering in the remaining stages of the transmitter and receiver analog chain and noise-less operation. Using (9), it can be seen that

$$\tilde{z}(t) = \left(C_1\tilde{s}(t) + C_2^*\tilde{s}^*(t)\right)e^{j(\Delta\omega t + \theta)} \quad (17)$$

where $\Delta\omega$ and θ account for the overall carrier frequency and phase offset between the transmitter and the receiver. Thus, we have

$$\tilde{r}(t) = K_1C_1\tilde{s}(t)e^{j(\Delta\omega t+\theta)} \quad (18a)$$

$$+ K_1C_2^*\tilde{s}^*(t)e^{j(\Delta\omega t+\theta)} \quad (18b)$$

$$+ K_2C_1^*\tilde{s}^*(t)e^{-j(\Delta\omega t+\theta)} \quad (18c)$$

$$+ K_2C_2\tilde{s}(t)e^{-j(\Delta\omega t+\theta)} \quad (18d)$$

The desired term in (18) is the one in the first line and the

terms in the second to fourth lines represent undesired images at the receiver due to transmitter and receiver IQ imbalances.

When $\Delta\omega = 0$, (18) simplifies to

$$\tilde{r}(t) = J_1\tilde{s}(t) + J_2\tilde{s}^*(t) \tag{19}$$

where J_1 and J_2 are constants given by

$$J_1 = K_1C_1e^{j\theta} + K_2C_2e^{-j\theta} \tag{20a}$$

$$J_2 = K_1C_2^*e^{j\theta} + K_2C_1^*e^{-j\theta}. \tag{20b}$$

Comparing (19) with (14), it can be concluded that when $\Delta\omega = 0$, the effect observed at the output of the receiver's IQ demodulator due to the IQ imbalance introduced at the transmitter is the same as the one due to an IQ imbalance introduced by the IQ demodulator. In a real transmission system there will be some carrier frequency offset between the transmitter and the receiver. However, the former observation suggests that the IQ imbalance introduced at the transmitter may be addressed after carrier frequency recovery using approaches designed to address the IQ imbalance introduced at the receiver.

B. Imbalance compensation

1) Tx IQ Imbalance Compensation: Using (1) and (3) with $\tilde{s}'(t) = s'_I(t) + js'_Q(t)$ denoting the equivalent baseband signal of the Tx IQ distorted signal $s(t)$ with respect to the transmitter carrier frequency ω_{tx}, we can derive the following matrix equation for Tx IQ imbalance distortion:

$$\begin{bmatrix} s'_I(t) \\ s'_Q(t) \end{bmatrix} = \begin{bmatrix} 1 & -g_{tx}\sin(\phi_{tx}) \\ 0 & g_{tx}\cos(\phi_{tx}) \end{bmatrix} \begin{bmatrix} s_I(t) \\ s_Q(t) \end{bmatrix}. \tag{21}$$

If the matrix in (21) is invertible ($g_{tx} \neq 0$ and $\phi_{tx} \neq \pm\pi/2$), which is the case in practical cases, the Tx NFS IQ compensation can ideally be achieved by performing a digital predistortion based on the inverse operation. In this case, we feed the IQ modulator with the predistorted signal $\varsigma(t) = \varsigma_I(t) + j\varsigma_Q(t)$, which is obtained as

$$\begin{bmatrix} \varsigma_I(t) \\ \varsigma_Q(t) \end{bmatrix} = \begin{bmatrix} 1 & \tan(\phi_{tx}) \\ 0 & 1/(g_{tx}\cos(\phi_{tx})) \end{bmatrix} \begin{bmatrix} s_I(t) \\ s_Q(t) \end{bmatrix}. \tag{22}$$

Compensation by inverse transformation requires knowledge of the gain and phase imbalance values, g_{tx} and ϕ_{rx}.

Techniques for compensating the IQ imbalance at the transmitter that have been proposed in the literature use tones as test signals (e.g., [10], [11], [12], [13]) or even from random data (e.g., [11], [14]). Compensating by using test tones is very powerful and can be used for initial calibration. During normal full-duplex operation of the transceiver, IQ imbalance compensation from the random transmitted data would be preferred. All these techniques of compensation require additional circuitry (including an extra ADC) at the transmitter to feed back measurements performed in the analog front-end.

Under certain conditions the Tx IQ imbalance can also be compensated in the receiver (see [9]). For this to be possible, the spectral images caused by the Tx IQ imbalance have to be emitted and a carrier frequency offset between transmitter and receiver is required to be able to decouple the Tx and Rx IQ imbalance effects in the receiver [15], [16]. As these conditions are fulfilled in the analyzed system, receiver-based compensation for both frequency-independent and frequency-selective Tx IQ imbalance is the approach investigated in this article, because it does not require any additional circuitry in the analog front-end of the transceiver and all the compensation can be performed by digital signal processing.

2) Rx IQ Imbalance Compensation: The down-converted complex-valued base-band signal $\tilde{r}(t)$ can be written as a function of $\tilde{z}(t)$ using the following matrix equation

$$\begin{bmatrix} r_I(t) \\ r_Q(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -g_{rx}\sin(\phi_{rx}) & g_{rx}\cos(\phi_{rx}) \end{bmatrix} \begin{bmatrix} z_I(t) \\ z_Q(t) \end{bmatrix}. \tag{23}$$

In case the matrix in (23) is invertible ($g_{rx} \neq 0$ and $\phi_{rx} \neq \pm\pi/2$), which is the case in practical cases, the Rx IQ imbalance can be ideally compensated by performing the inverse operation

$$\begin{bmatrix} \alpha_I(t) \\ \alpha_Q(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \tan(\phi_{rx}) & 1/(g_{rx}\cos(\phi_{rx})) \end{bmatrix} \begin{bmatrix} z'_I(t) \\ z'_Q(t) \end{bmatrix}, \tag{24}$$

where $\alpha(t) = \alpha_I(t) + j\alpha_Q(t)$ is the output of the Rx IQ imbalance compensator and ideally would yield the desired baseband signal $\tilde{z}(t)$. Compensation by inverse transformation requires knowledge of the gain and phase imbalance values g_{rx} and ϕ_{rx}, which can be derived by using statistics and the correlation properties of the I and Q signals as proposed in [17].

IV. NFS IQ Imbalance Simulation Results

A. System Model

The system model depicted in Fig. 3 constitutes the basis for the simulations of the impact of Tx and Rx IQ imbalance impairments and their respective compensation algorithms on the performance of the E-band transceiver.

Random data information is generated as a sequence of IQ symbols, *D*, by using a 64QAM mapper. *D* is then filtered through an appropriate root-raised-cosine (RRC) filter to create a pulse-shaped base-band signal, $x(t)$. The Tx Baseband Filtering and Mixing models the data processing as shown in Fig. 1. There is the option of employing Tx IQ predistortion to mitigate the NFS Tx IQ imbalance, but for the purposes of this paper, this block will be deactivated.

Considering an additive white Gaussian noise (AWGN) channel, the signal $\tilde{z}$ at the output of the channel model is given by:

$$\tilde{z}(t) = \tilde{s}(t) \cdot \mathrm{e}^{j2\pi ft} + n(t), \tag{25}$$

where $\tilde{z}(t)$ is the low-pass equivalent of the transmitted signal, $\tilde{s}(t)$ is the transmitted signal and $\mathrm{e}^{j2\pi ft}$ represents the carrier frequency offset (CFO), with f as the frequency offset. This CFO is produced by the difference between the oscillators of the transmitter and receiver. Finally, $n(t)$ corresponds to a complex-valued white Gaussian noise process.

On the receiver side, the Rx Baseband Filtering and Mixing models the receiver analog front-end structure shown in Fig. 1, including Rx IQ imbalance. NFS IQ Imbalance Compensation

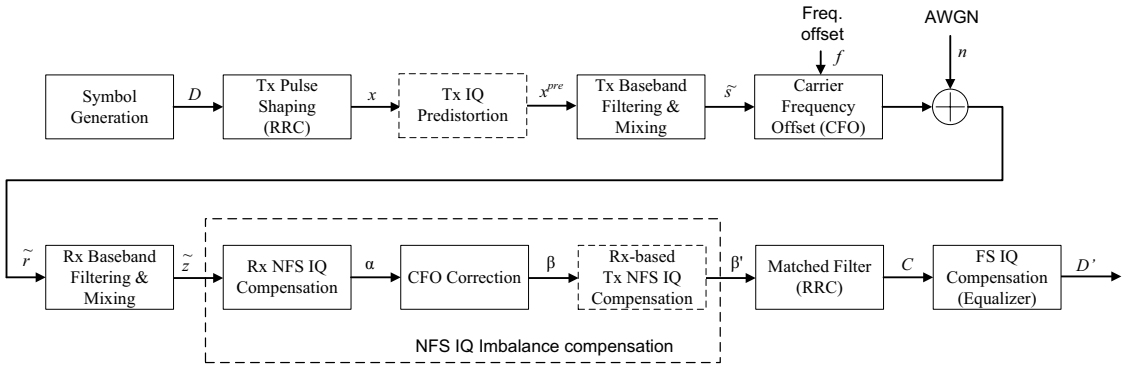


Fig. 3. System Model for IQ Imbalance Simulations.

is performed prior to matched filtering by following a multi-stage approach that compensates for both Tx and Rx IQ Imbalance:

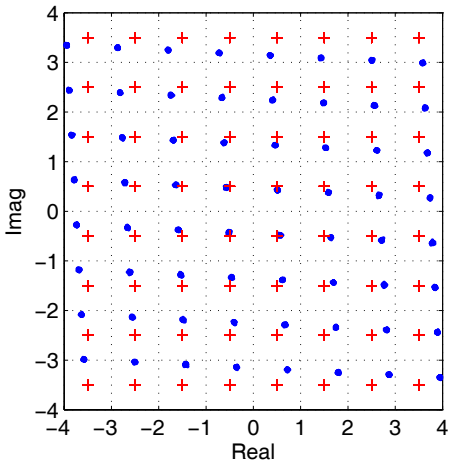
1) The first stage compensates for the Rx NFS IQ imbalance based on the algorithms in Section III-B.
2) A CFO correction stage ideally compensates the CFO introduced by the difference of the transmitter and receiver, using the relation $\beta(t) = \alpha(t) \cdot \mathrm{e}^{-i2\pi ft}$
3) A third stage performs Tx NFS IQ imbalance compensation prior to the matched ltering by re-applying the algorithms in Section III-B .

After NFS IQ Imbalance Compensation the signal is filtered with a matched RRC filter. The equalizer used to compensate the FS IQ imbalance is ignored in this section (i.e., $D' = C$). The received symbols *D'* are compared with the transmitted symbols *D* to quantify the performance of the whole system.

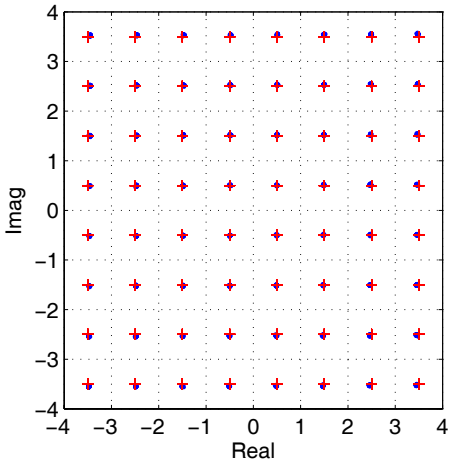
B. NFS IQ Imbalance results

Fig. 4(a) illustrates the effect of NFS IQ imbalance. For illustration purposes, a noiseless transmission is considered. Crosses correspond to the constellation where perfect transmission takes places, and the dots correspond to a transmission with IQ imbalances both at transmitter and receiver, assuming zero carrier frequency offset. The IQ imbalances considered were 0.5 dB and 3 degrees in gain and phase, respectively, in the transmitter and 1 dB and 3 degrees in gain and phase, respectively, in the receiver. The signs of the imbalances were selected in the transmitter and the receiver so that they combine in the worst possible distortion. Fig. 4(b) shows the constellation at the receiver when the NFS IQ imbalance compensation is active. It can be seen that the NFS IQ imbalance compensation approach is able to remove the distortion effect on the constellation.

Fig. 5 shows the impact of the NFS IQ imbalance in the performance of the transceiver. The curve labeled *'without NFS IQ'* is the performance of the transceiver when there is no IQ imbalance. The curve labeled *'NFS IQ'* is the performance of the transceiver with IQ imbalance at both transmitter and receiver. The curve labeled *'NFS IQ with comp'* is the



(a) Without compensation



(b) With compensation

Fig. 4. NFS IQ imbalance

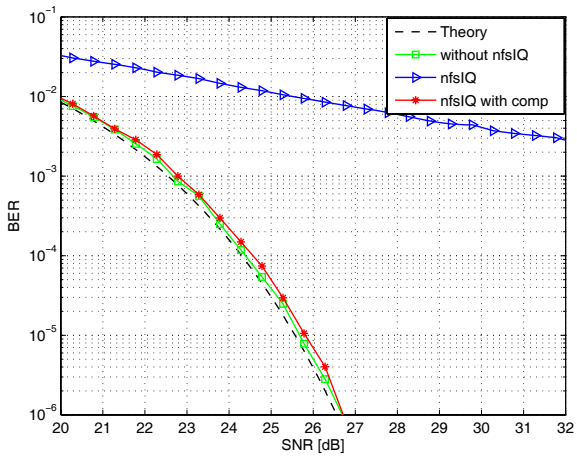


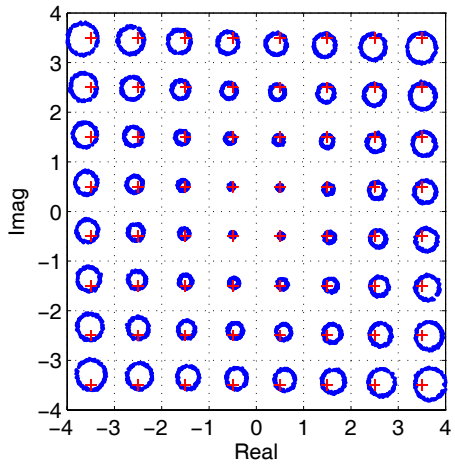
Fig. 5. System performance in the presence of NFS IQ imbalance.

performance when the IQ imbalance compensation is active. The NFS IQ imbalance compensation is able to reduce the losses to a few tenths of a dB.

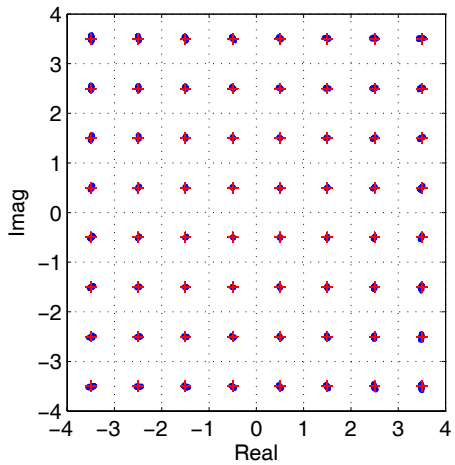
Fig. 6(a) illustrates the effect of NFS IQ imbalance when there is a residual carrier frequency offset equal to $\Delta\omega = 2\pi \cdot 5 \cdot 10^{-6}/T$ after the CFO correction in Fig. 3, where T is the symbol period. For proper symbol detection and Bit Error Rate(BER) estimation, the residual carrier frequency offset was compensated in the simulations at the input of the receiver's matched-filter. The figure shows the corrected constellations after the final residual carrier frequency offset compensation. The transmitter's IQ imbalance manifests itself as rotations of the constellation around the distorted constellation due to the receiver's IQ imbalance.

Fig. 6(b) shows the constellation when the IQ compensation algorithms are active for the same residual carrier frequency offset. It can be seen that the distortion of the constellation has been reduced despite some residual carrier frequency offset.

Fig. 7 shows the impact of the residual carrier frequency offset on the performance of the NFS IQ imbalance compensation approach presented in Section III-B. The different curves correspond to different values of the normalized residual carrier frequency offset $\Delta\omega T/(2\pi)$. It can be seen that a very accurate carrier frequency offset correction is needed for proper compensation of the transmitter IQ imbalance with residual CFO below $\Delta\omega T/(2\pi) = 5 \cdot 10^{-6}$, which can be achieved via application of state of the art coarse and fine frequency synchronization algorithms [18].



(a) Without compensation



(b) With compensation

Fig. 6. NFS IQ imbalance with normalized residual carrier frequency offset $\Delta\omega T/(2\pi) = 5 \cdot 10^{-6}$.

V. Frequency-selective IQ Imbalance

A. Imbalance analysis and models

The above analysis considers that the imbalance is constant with frequency, and thus we talk about non-frequency-selective IQ imbalance. Nevertheless, the elements placed in the I-datapath and Q-datapath present dispersion in their parameters. Therefore, there will always be a difference between the performance of the components placed in the I-datapath and the components placed in the Q-datapath. These dispersions can contribute to non-frequency-selective distortion, and likewise to impairment due to the LOs, or to frequency-selective distortion. For instance, Fig. 8 compares the frequency responses (magnitude and phase) of the designed passive baseband filters by considering typical values for the fabrication tolerances of their components. L0C0 is the nominal case. LpCp is the corner where inductors and capacitances have their maximum value. LmCm is the corner where inductors and capacitors have their minimum value. Fig. 9 shows the difference between the two extreme corners. This kind of mismatch between the components in the I and Q datapaths is a frequency-selective IQ imbalance.

Fig. 10 shows the BER performance of the system obtained from the model that compensates the non-frequency-selective IQ imbalance, but does not correct the frequency-selective mismatch. The FS IQ imbalance is considered at both the transmitter and the receiver sides of the transceiver. The curve labeled *'IbbfiltL0C0, QbbfiltL0C0'* represents the nominal

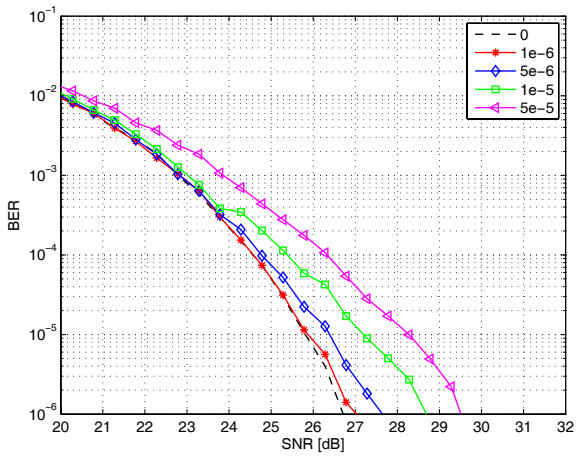
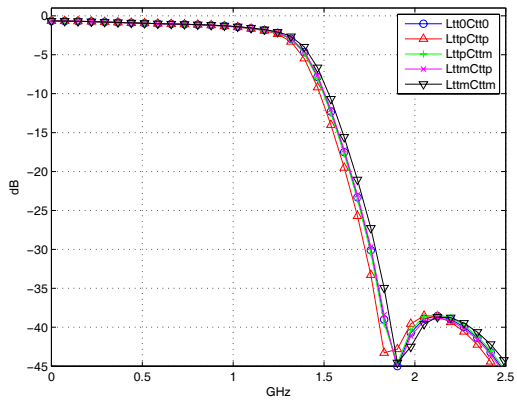
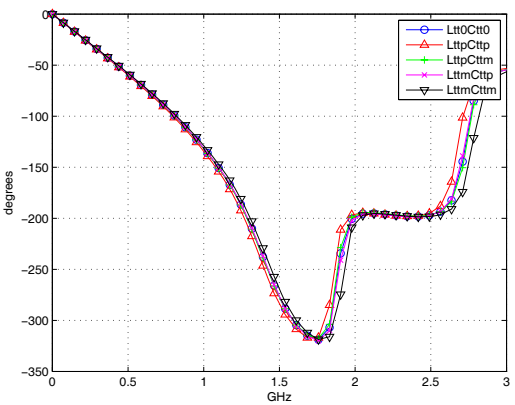


Fig. 7. Performance of NFS IQ imbalance compensation approach for different values of $\Delta\omega T/(2\pi)$.

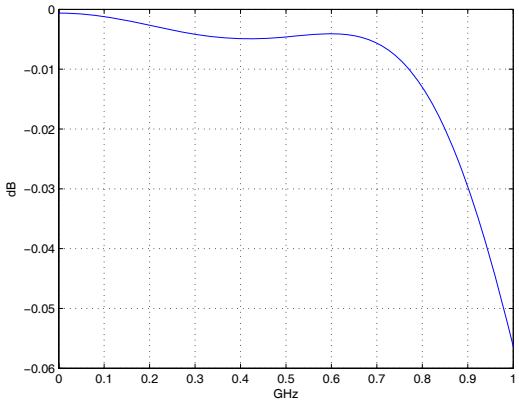


(a) Magnitude variation

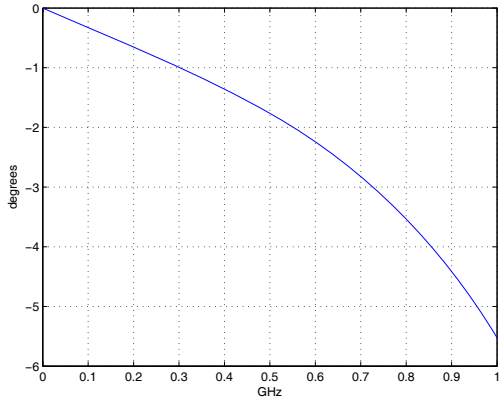


(b) Phase variation

Fig. 8. Variation of frequency response of base-band filters.



(a) Magnitude variation



(b) Phase variation

Fig. 9. Frequency response of base-band filters in extreme corners.

case, i.e., the case without FS IQ imbalance. The curve labeled *'IbbfiltLpCp, QbbfiltLmCm'* is the case in which an LpCp base-band filter has been considered in the I-datapath and an LmCm base-band filter has been chosen for the Q-datapath in the transmitter and the receiver. From the BER performance it can be concluded that the loss due to FS IQ imbalance significantly degrades the performance of the system, since for a BER of 10^{-6} in the case of uncoded 64QAM, a loss of more than 2dB is shown.

In the literature more articles focusing on the modeling and compensation of non-frequency-selective IQ imbalance can be found. However, this approach has become insufficient as the use of wider bandwidth signals has become more prevalent. Therefore, as we can conclude from Fig. 10, for wide bandwidth signals the above technique for compensating for the non-frequency-selective IQ imbalance is generally insufficient, as the IQ imbalance is likely to be frequency-dependent.

The frequency-dependent imbalances of the quadrature mixing front-end may include contributions from the DAC/ADC, low-pass filters, as well as the signal paths themselves. To formally analyze the detrimental effect of these IQ mismatches,

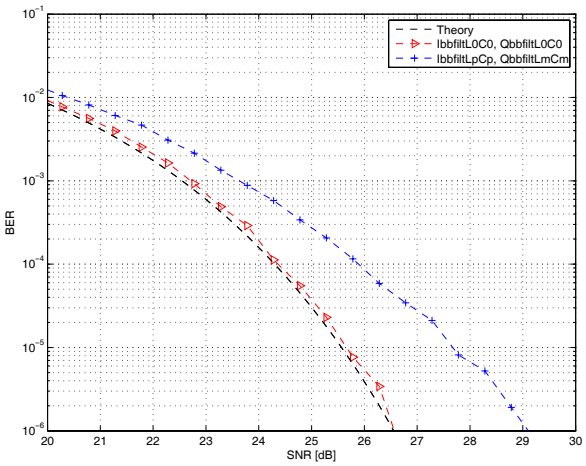


Fig. 10. System performance in the presence of FS IQ imbalance..

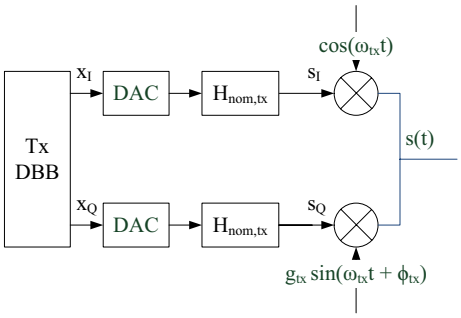


Fig. 11. Tx IQ Imbalance Model.

the combined effect can simply be modeled as a pair of two imbalanced base-band filters. The models depicted in Fig. 11 and Fig. 12 have been devised as a basis for the development of algorithms for Tx and Rx IQ imbalance compensation.

For Tx IQ imbalance in Fig. 11 the filters $H_{nom,tx}(f)$ are assumed to be ideal low-pass filters with $H_{nom,tx}(f) = 1$ for $|f| \leq B_{tx}/2$ and $H_{nom,tx}(f) = 0$ for $|f| > B_{tx}/2$, where B_{tx} is the bandwidth of the transmitted signal of interest. In analyzing NFS IQ imbalance, these filters are assumed to be virtually ineffective. The filters $H_{I,tx}(f)$ and $H_{Q,tx}(f)$ are modeling frequency-dependent or frequency-selective mismatches between the I and Q path, as shown in Fig. 8 and Fig. 9.

The corresponding model for Rx IQ imbalance is depicted in Fig. 12 (see also [19], [6]). The filters $H_{nom,rx}(f)$ are assumed to be ideal low-pass filters with $H_{nom,rx}(f) = 1$ for $|f| \leq B_{rx}/2$ and $H_{nom,rx}(f) = 0$ for $|f| > B_{rx}/2$, where B_{rx} is the bandwidth of the received signal of interest. Similar to the TX IQ imbalance model in Fig. 11, the filters $H_{I,rx}(f)$ and $H_{Q,rx}(f)$ are modeling frequency-dependent or frequency-selective mismatches between the I and Q paths.

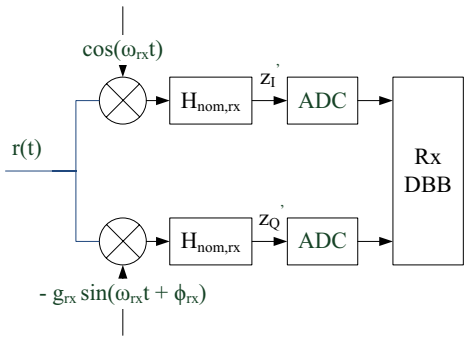


Fig. 12. Rx IQ Imbalance Model.

B. Frequency-Selective IQ Imbalance Compensation

Methods for compensating Rx FS IQ imbalance are reported in [6], [5]. In [6], an adaptive filter is employed to perform interference cancellation. In [5], a simplified compensation method, based on mean values of the FS IQ imbalance, is presented. However, this method assumes prior knowledge of the mean values over frequency for the frequency-selective part of the amplitude and phase imbalances.

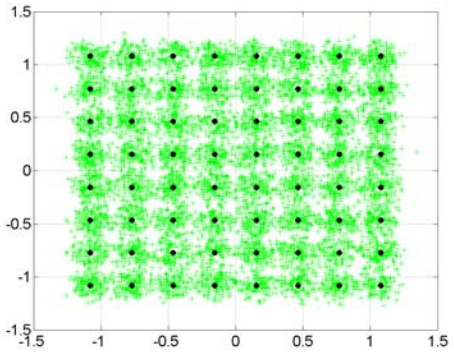
In this article, a linear adaptive equalizer (LAE) is deployed for the FS IQ imbalance compensation as shown in the system model in Fig. 3. In addition to NFS IQ Imbalance compensation analyzed in Section IV-B, the LAE is activated as the final stage in performing compensation of FS IQ imbalance after matched filtering. The LAE performs linear filtering of the Matched Filter output as defined by (26). The LAE facilitates FS IQ compensation without prior knowledge of the frequency responses of frequency-selective amplitude and phase imbalances. The L coefficients w_i in (26) are calculated and adapted based on known symbol sequences in the data stream using the Least Mean Square (LMS) algorithm. For the following simulations an equalizer length of $L = 15$ is used.

$$D'[k] = \sum_{i=0}^{L-1} w_i C[k-i]. \tag{26}$$

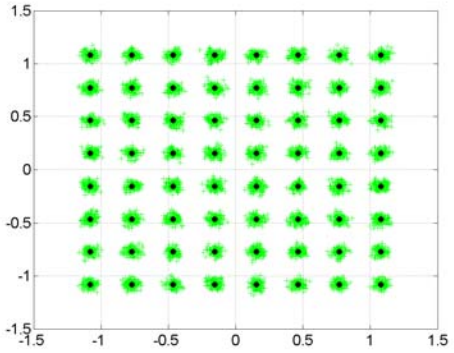
The system performance in terms of FS IQ Imbalance distortion is evaluated in terms of residual carrier-to-distortion ratio after the LAE is defined by the relation

$$\frac{C}{ISI} = \frac{E\left\{|D|^2\right\}}{E\left\{|D-D'|^2\right\}}, \tag{27}$$

where $E\left\{|D|^2\right\}$ is the energy of the transmitted signal D and $E\left\{|D-D'|^2\right\}$ denotes the energy of the received signal error, i.e., distortion caused by residual inter-symbol-interference (ISI) and image interference. Simulation results are provided in Section VI.



(a) Without equalizer



(b) With equalizer

Fig. 13. Signal Constellations (64QAM)

VI. FS IQ Imbalance Simulation Results

Fig. 13 shows the constellation diagrams for 64QAM after the RRC matched filter and the equalizer, respectively. The circles in the diagrams correspond to the ideal constellation points. When comparing Fig. 13(a) with Fig. 13(b) the reduction in residual distortion achieved by the equalizer can be observed.

For a BER of 10^{-6} in the case of uncoded 64QAM a $C/(I+ISI) = 26.54\,dB$ is required at the input to the demapper, which corresponds to the output of equalizer *D'*. Table I summarizes the performance for different transmit and receive filter pairs based on the frequency responses shown in Fig. 8 and Fig. 9. Simulations 1 to 3 have been run with identical filters for both quadrature components in the transmitter and receiver. The residual ISI is in the range of 38 dB. Simulations 4 to 9 have been run with imbalanced quadrature filters. The FS IQ imbalance, i.e., the difference between frequency responses, is most severe in simulations 4 and 5, where the corner cases LpCp and LmCm for the tolerances concerning the frequency responses have been used. The residual ISI after the equalizer is reduced below levels where 64QAM can be decoded. For less severe filter imbalances, as in simulations 6 to 9, the residual ISI is about 4 to 5 dB better.

TABLE I
FS IQ Imb. Performance

	$H_{I,tx}(f)$	$H_{Q,tx}(f)$	$H_{I,rx}(f)$	$H_{Q,rx}(f)$	C/ISI [dB]
1	L0C0	L0C0	L0C0	L0C0	38.5
2	LpCp	LpCp	LpCp	LpCp	38.4
3	LmCm	LmCm	LmCm	LmCm	37.8
4	LpCp	LmCm	LpCp	LmCm	27.1
5	LmCm	LpCp	LmCm	LpCp	29.0
6	LpCp	L0C0	LpCp	L0C0	31.9
7	L0C0	LpCp	L0C0	LpCp	34.6
8	LmCm	L0C0	LmCm	L0C0	34.7
9	L0C0	LmCm	L0C0	LmCm	34.6

VII. Conclusion

In this article, a theoretical analysis of the IQ imbalance in a heterodyne transceiver with zero-second-IF has been presented. Digital signal processing at the receiver has been evaluated based on a multi-stage approach for the compensation of both transmitter and receiver IQ imbalance, as well as carrier frequency offset.

Simulation results were presented for a zero-second-IF transceiver using 64-QAM with a signal bandwidth of 2GHz. For NFS IQ imbalance, the simulation results show that receiver-based Tx IQ imbalance compensation can be achieved with negligible degradation in overall system performance. This is achieved when accurate frequency synchronization is performed in the receiver such that the residual carrier frequency offset is reduced below specified limits. For FS IQ imbalance, simulation results have demonstrated that FS IQ imbalance can be mitigated by a receiver-based compensation concept based on linear equalization.

Acknowledgment

The research leading to these results has received funding from the European Community's FP7/2007-2013 Framework Programme under grant agreement no. 317957. Consortium: CEIT, Fraunhofer IIS, Alcatel-Lucent, CEA-Leti, IXYS, Silicon Radar, ST, Sivers IMA, OTE.

References

[1] A. Rezola, J. Sevillano, R. Berenguer, I.Velez, M. Leyh, M. Lorenzo, and A. Vargas, "Non-frequency-selective i/q imbalance in zero-if transceivers for wide-band mmw links," in *The Tenth International Conference on Wireless and Mobile Communications ICWMC*, 2014, pp. 136–141.

[2] *Fixed Radio Systems; Characteristics and requirements for point-to-point equipment and antennas; Part 1: Overview and system-independent common characteristics*, ETSI EN 302 217-1, Sept. 2012.

[3] *Fixed Radio Systems; Characteristics and requirements for point-to-point equipment and antennas; Part 2-2: Digital systems operating in frequency bands where frequency co-ordination is applied; Harmonized EN covering the essential requirements of article 3.2 of the R&TTE Directive*, ETSI EN 302 217-2-2, Sept. 2012.

[4] G. Fettweis, M. Lohning, D. Petrovic, M. Windisch, P. Zillmann, and W. Rave, "Dirty RF: a new paradigm," in *Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on*, vol. 4, 2005, pp. 2347–2355 Vol. 4.

[5] M. Mailand, R. Richter, and H.-J. Jentschel, "IQ-imbalance and its compensation for non-ideal analog receivers comprising frequency-selective components," *Advances in Radio Science*, vol. 4, pp. 189–195, Sept. 2006.

[6] M. Valkama, M. Renfors, and V. Koivunen, "Compensation of frequency-selective i/q imbalances in wideband receivers: models and algorithms," in *Wireless Communications, 2001. (SPAWC '01). 2001 IEEE Third Workshop on Signal Processing Advances in*, 2001, pp. 42–45.

[7] V. Dyadyuk, J. Bunton, J. Pathikulangara, R. Kendall, O. Sevimli, L. Stokes, and D. Abbott, "A multigigabit millimeter-wave communication system with improved spectral efficiency," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 55, no. 12, pp. 2813–2821, Dec 2007.

[8] B. Razavi, "Design considerations for direct-conversion receivers," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 44, no. 6, pp. 428–435, Jun 1997.

[9] L. Antilla, "Digital Front-End Processing with Widely-Linear Signal Models in Radio Devices," Ph.D. dissertation, Tampere University of Technology, 2011.

[10] J. Cavers and M. Liao, "Adaptive compensation for imbalance and offset losses in direct conversion transceivers," *Vehicular Technology, IEEE Transactions on*, vol. 42, no. 4, pp. 581–588, Nov 1993.

[11] J. Cavers, "New methods for adaptation of quadrature modulators and demodulators in amplifier linearization circuits," *Vehicular Technology, IEEE Transactions on*, vol. 46, no. 3, pp. 707–716, Aug 1997.

[12] A. Nassery, S. Byregowda, S. Ozev, M. Verhelst, and M. Slamani, "Built-in-self test of transmitter i/q mismatch using self-mixing envelope detector," in *VLSI Test Symposium (VTS), 2012 IEEE 30th*, 2012, pp. 56–61.

[13] S. D'souza, F. Hsiao, A. Tang, S.-W. Tam, R. Berenguer, and M.-C. Chang, "A 10-bit 2-gs/s dac-ddfs-iq-controller baseband enabling a self-healing 60-ghz radio-on-chip," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 60, no. 8, pp. 457–461, Aug 2013.

[14] X. Huang and M. Caron, "Efficient transmitter self-calibration and amplifier linearization techniques," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, 2007, pp. 265–268.

[15] P. Rykaczewski and F. Jondral, "Blind i/q imbalance compensation in multipath environments," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, May 2007, pp. 29–32.

[16] J. de Witt and G.-J. van Rooyen, "A blind i/q imbalance compensation technique for direct-conversion digital radio transceivers," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 4, pp. 2077–2082, May 2009.

[17] H.-J. Jentschel, "Direct conversion receivers - expectations and experiences," in *RF Front End Architectures, IEEE MTT-S 2000, Boston, Workshop*, June 2000.

[18] H. Meyr, M. Moeneclaey, and S. A. Fechtel, *Digital Communication Receivers*. Wiley, 1998.

[19] M. Valkama, M. Renfors, and V. Koivunen, "Advanced methods for i/q imbalance compensation in communication receivers," *Signal Processing, IEEE Transactions on*, vol. 49, no. 10, pp. 2335–2344, Oct 2001.

Designing Towards A Fully Monolithic Envelope-Tracking SiGe Power Amplifier for Broadband Wireless Applications

Yan Li [1, 2], Jerry Lopez [1,3], and Donald Y.C. Lie [1]
[1] Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX, USA
[2] Qorvo Inc., Phoenix, AZ, USA
[3] NoiseFigure Research Inc., Lubbock, TX, USA
yan.li.ttu@gmail.com

Abstract— **This paper presents an overall design path towards our fully monolithic envelope-tracking (ET) power amplifier (PA) in a 0.35 µm SiGe BiCMOS technology. First, a discrete hybrid switching supply modulator (SM) is designed to study the effects of its switching frequency and bandwidth limitation to the linearity and efficiency of an overall ET-PA. Next, the same circuit configuration is used for our CMOS SM which modulates our self-biased cascode SiGe PA with the proposed envelope shaping function. By analyzing the power losses of the CMSO SM under our special ET operation, we show the SM can still achieve high efficiency with a relatively low switching frequency, which helps to improve the system linearity. At 1.9 GHz, our BiCMOS cascode ET-PA delivers the maximum linear output power (P_{out}) of 24/23.4 dBm with the overall power-added-efficiency (PAE) of 41%/38% for the long-term evolution (LTE) 16QAM 5/10 MHz signals, respectively. Meanwhile, good linearity and low spurious emission are achieved without the need of digital predistortion (DPD).**

Keywords- Broadband; envelope-tracking (ET); hybrid switching supply modulator (SM); long-term evolution (LTE); SiGe; power amplifier (PA); WiMAX

I. INTRODUCTION

Broadband 3G/4G/WLAN wireless standards utilize highly spectral-efficient modulation schemes with inherent non-constant envelope signals having high peak-to-average power ratios (PARs). In a typical mobile transmitter (TX), the radio frequency (RF) power amplifier (PA) tends to be the most power-hungry block. Although high efficiency can be achieved by the PA at the saturated output power (P_{sat}) for GSM handset systems, its high distortion at P_{sat} violates the linearity specs for non-constant envelope signals (e.g., in the case of long-term evolution (LTE)). Therefore, the PA is usually required to be backed off from its P_{sat} for non-constant envelope signals, resulting in poor efficiency.

The envelope tracking (ET) technique is a promising solution to improve the PA efficiency for broadband applications [1]-[12]. In ET-PA systems, the envelope signal from a supply modulator (SM) dynamically modulates the PA collector or drain voltage in response to the instantaneous output power (P_{out}). In practical implementation, a hybrid switching SM (also called as a linear-assisted switching SM) is often used to achieve a good balance between the efficiency and the wideband envelope tracking [1]-[16]. Although many articles have reported excellent efficiency by using this hybrid switching SM, its design details still need to be studied and optimized with a specific PA to achieve the best trade-off between the efficiency and linearity for the overall ET-PA. In ET systems, both the PA and SM are the distortion sources, thus if the SM can be tuned for better linearity without suffering a great efficiency reduction, the PA can be pushed into deeper compression, leading to higher overall efficiency. Moreover, a hybrid switching SM is typically designed with a high average switching frequency (e.g., ≥ 2 MHz in [9]-[11]), creating out-of-band spurs which may violate the spectral mask of interest and spurious emission specs. To sufficiently suppress the high frequency switching ripples, the linear stage of the hybrid switching SM must have an ultra-wide bandwidth, inevitably increasing the design complexity [17]. Therefore, it becomes necessary to understand how much the average switching frequency can be lowered without hurting much on the SM efficiency. To address the concerns mentioned above, in this paper, we will present our complete design path towards a fully monolithic ET-PA in a silicon based technology.

In Section II, a discrete hybrid switching SM is designed to study the effects of the bandwidth limitation and switching frequency to the linearity and efficiency of the overall ET-PA system. The discrete SM is paired with a SiGe PA in Section III to compare the performances between the stand-alone PA (i.e., the fixed supply PA) and the ET-PA. The measurement data verifies that our design approach of the SM is suitable for high PAR wideband applications.

As the expansion to the work [1], we will present our fully monolithic BiCMOS ET-PA in Section IV. A CMOS SM will be integrated with the cascode SiGe PA in a 0.35 µm SiGe BiCMOS technology to form a fully monolithic ET-PA system. In contrast to the discrete solution described in Section II, the single-chip ET-PA integration solution has several benefits: (1) it reduces cost; (2) it has an integrated signal path providing better signal integrity [11], and (3) from the design perspective, the PA and the SM can be optimized together to achieve better linearity and overall efficiency. Unlike common emitter PAs, our cascode PA structure relieves the voltage stress on the power transistor, while also requires some special envelope shaping function to improve its linearity at the ET operation. In Section IV, we will first highlight the self-biasing structure of our cascode PA design and our proposed envelope shaping method based on [3]. Moreover, we will analyze the power

losses of the CMOS SM in details in responding to our special cascode ET-PA operation. Although the same SM was presented in [3], [4], this paper further demonstrates how our SM can achieve high efficiency with a relatively low average switching frequency, which helps to improve the overall ET linearity. Based on the measurement data in Section V, our BiCMOS cascode ET-PA achieves state-of-the-art overall efficiency with good linearity performance for LTE 16QAM 5/10 MHz signals. No digital predistortion (DPD) technique will be used in our measurement.

II. Design Insights for Envelope Tracking

A. Design of Common-Emitter SiGe Power Amplifier

A monolithic 1-stage common-emitter SiGe PA is used here as an example to form an ET-PA system to study the trade-offs for the hybrid switching SM design. This PA was designed and fabricated in the IBM 7HP 0.18µm SiGe BiCMOS technology [5], [6]. The simplified schematic and die picture of the PA are shown in Fig. 1. The high-breakdown heterojunction bipolar transistor (HBT) option is used for the PA design with a total emitter-area of 220 µm^2 (typical BV_{CEO}= 4.2 V, BV_{CBO}= 12.5 V). This monolithic SiGe PA was tested on an FR4 PCB. To achieve high power-added-efficiency (PAE), the RF choke (RFC) inductor was left off-chip for high Q at 2.4 GHz. Additionally, the output tank inductor was realized by a bondwire, and more than 4 downbonds (i.e., bondwires at the emitter node) are used to reduce the grounding parasitic inductance [5]. No other off-chip elements are needed for the input and output matching.

It is important to characterize the PA thoroughly before designing the SM for optimal ET-PA performances, as the collector impedance presented by the PA (R_{load}) will affect the efficiency and linearity performance of the SM. Fig. 2 shows the measured PAE vs. output power (P_{out}) at different supply voltage V_{CC} in the continuous wave (CW) mode. For the fixed-supply PA, its PAE reduces rapidly when P_{out} drops, while the PAE at low P_{out} can be greatly enhanced by varying V_{CC} as shown by the dash curve. The dash curve is the ideal operating trajectory of an ET-PA by tracking the peak PAE points at different P_{out} levels. Fig. 2 also plots the R_{load} presented by the PA to the SM, which is calculated from the DC supply voltage and the measured DC supply current of the PA at each peak PAE point. The R_{load} changes roughly from 70 Ω to 10 Ω at P_{out} from 8 dBm to 20 dBm.

B. Hybrid Switching Supply Modulator

A proper SM design is critical to achieve the best overall efficiency and linearity performances for an ET-PA. As reported in [2], the finite bandwidth and the associated group delay of the SM are large contributors of nonlinearity in an ET-PA. In addition, to take advantage of the efficiency enhancement provided by the ET technique, the SM needs to maintain high efficiency throughout the ET-PA operation. The overall efficiency of an ET-PA system ($\eta_{ET\text{-}PA}$) is the product of the SM efficiency (η_{SM}) and the PA collector efficiency ($\eta_{PA,CE}$), which is expressed as:

$$\eta_{ET-PA} = \eta_{SM} \cdot \eta_{PA,CE} \quad (1)$$

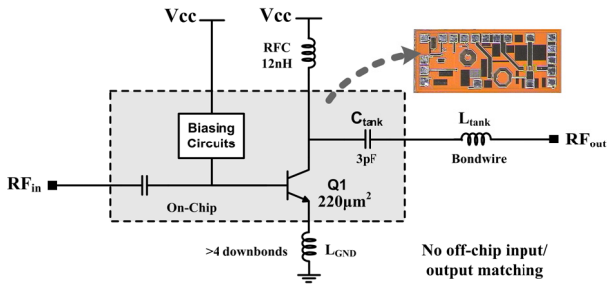


Fig. 1. Simplified schematic and die picture of the 1-stage PA designed and fabricated in IBM 7HP 0.18 µm SiGe BiCMOS technology

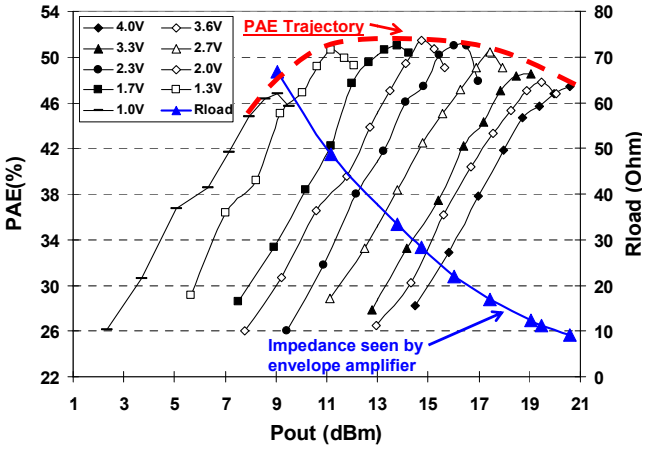


Fig. 2. Measured PAE vs. P_{out} of the SiGe PA in the CW mode, and the R_{load} presented by the PA to the SM

Therefore, the design goals of an SM are high efficiency and wide bandwidth to track the instantaneous input envelope.

1) Hybrid Switching Structure

The envelope signal is extracted from the modulated I/Q (i.e., in-phase/quad-phase) signals from the LTE/WiMAX baseband and then feed into the SM. Such nonlinear transformation will expand the bandwidth of the envelope by a factor of 5-10 compared with the original signal bandwidth [2], [7]. Conventionally, the SM can be implemented in the form of a linear regulator (e.g., a low dropout regulator (LDO) as in [18]), as the linear topology offers wide bandwidth and can reduce much of the output ripples. Nonetheless, the power efficiency of linear regulator is very poor when the output voltage level is low [18], making it unsuitable for high PAR signals for 3G/4G applications. On the other hand, a switching regulator has high efficiency across a broad range of output voltage, but it produces significant output ripples and its bandwidth is constrained to be a fraction of the switching frequency [19], making it suitable only for narrowband applications such as the North American Digital Cellular (NADC) in [19]. With a rather high switching frequency, switching regulators could be pushed into high data-rate systems, but inevitably causing high switching loss that limits the efficiency (e.g., ~76% maximum efficiency for WCDMA in [20]). A high switching frequency may also degrade the ET-PA linearity

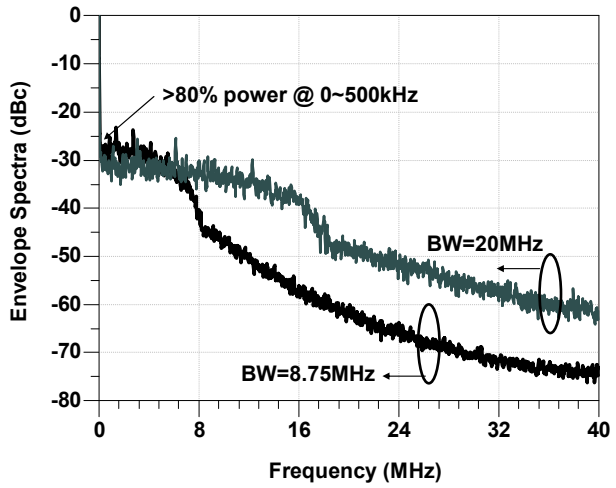


Fig. 3. Simulated envelope spectra of WiMAX 8.75/20 MHz signals

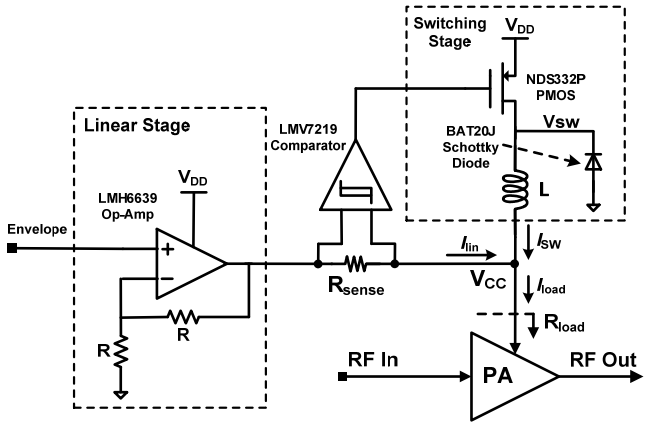


Fig. 4. Schematic of the discrete hybrid switching SM designed by using COTS components

considerably, which often defeats the purpose of using switching regulators for any supply-modulated PAs.

Many recent reports on the wideband SM design for ET-PA have combined the advantages of a wideband linear regulator and a high efficiency switching converter in various ways [1]-[17]. Fig. 3 shows the simulated envelope spectra of WiMAX 8.75 MHz and 20 MHz signals (PAR of ~10.2 dB). An important characteristic of the envelope spectrum is that ~80% of envelope power resides from DC to several kHz, while over 99% of the envelope power resides within DC to 8MHz for the 8.75 MHz signal, and within DC to 20 MHz for the 20 MHz signal, respectively. Such a characteristic of the envelope spectrum implies that a hybrid switching structure can achieve a good balance between efficiency and bandwidth tracking for an ET-PA system. The hybrid switching SM consists of a wideband but low efficiency linear stage and a high efficiency narrowband switching stage. The hybrid structure lessens the requirements of the switching stage, since the fast transients of the envelope signal are taken care of by the wideband linear stage, while the switching stage handles the DC and slow moving signals with high efficiency. The overall efficiency of the entire SM (η_{SM}) is a combination of the switching stage efficiency (η_{SW}) and the linear stage efficiency (η_{lin}), as expressed by:

$$\frac{1}{\eta_{SM}} = \frac{\alpha}{\eta_{SW}} + \frac{1-\alpha}{\eta_{lin}} , \quad (2)$$

where α is the ratio of the output power from the switching stage to the total output power of the SM [7], [16].

2) Discrete Supply Modulator Design

The hybrid switching SM is implemented by using commercial-of-the-shelf (COTS) components to investigate the overall efficiency and linearity trade-off for an ET-PA system. Fig. 4 shows the circuit implementation of the discrete SM using an operational amplifier (Op-Amp) as the linear stage, and a buck converter as the switching stage. The buck converter supplies the slow slew-rate load current (I_{SW}) that contributes to the majority of the load current (I_{load}) to ensure high efficiency, while the wideband linear Op-Amp stage operates in a feedback mode to track the high slew-rate current (I_{lin}). Additionally, the ripples caused by the buck converter will be attenuated and/or filtered by the linear Op-Amp. The smooth transition between the linear stage and the switching stage is realized by a hysteretic current feedback control. The hysteretic current feedback control consists of a current sensing resistor R_{sense} that senses the output current of the linear stage and a hysteresis comparator to control the buck converter. The value of the sensing resistor R_{sense} is chosen to be 1 Ω in this case, as it needs to be much smaller than R_{load} (i.e., the load impedance presented by the PA) to achieve high efficiency.

C. Efficient and Linearity of the Hybrid Switching SM

Although there are many reports on the efficiency of the SM in the literature [7]-[13], the effects of its bandwidth and switching frequency have not been studied as rigorously. Since the switching ripples (generated by high switching frequencies) and the bandwidth limitation are two major factors that cause distortions at the SM output, understanding their effects helps to optimize both efficiency and linearity of an overall ET-PA system. In this section, the nonlinearities of a discrete hybrid switching SM will be discussed.

1) Bandwidth of the Hybrid Switching SM

The linear stage (i.e., the Op-Amp) should have sufficient bandwidth not only to track the high frequency contents of the envelope signal with high fidelity as suggested by previous works [7]-[13], but also to suppress the switching ripples/noise generated from the switch converter. The switching ripples beyond the bandwidth of the linear stage can distort the envelope signal, and also be mixed with the modulated carrier in the PA to cause large spurious noise at the PA output, potentially degrading the system linearity.

To investigate the effect of the bandwidth, an Op-Amp behavior model provided by Agilent's ADS is used to replace the realistic Op-Amp model in the simulation, so that its bandwidth can be changed manually. The realistic SPICE models are still used for other blocks of the SM. Fig. 5 shows

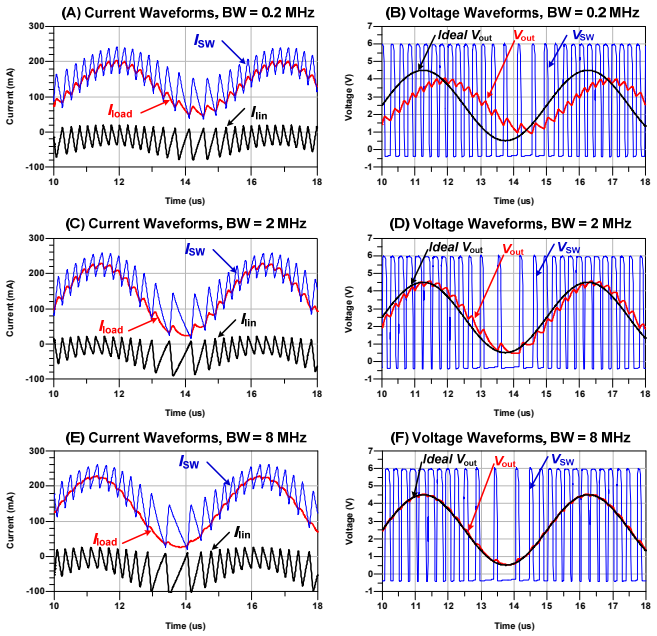


Fig. 5. Simulated (A, C, E) current and (B, D, F) voltage waveforms of the SM; the behavior model is used for the Op-Amp with different 1-dB bandwidths, while realistic SPICE models are used for other blocks of the SM. Input voltage = 1.25+$sin(2\pi \cdot 200\text{kHz} \cdot t)$ V, L = 4.7 µH, R_{load} = 20 Ω

the simulated current and voltage waveforms of the SM using different 1-dB bandwidths of the Op-Amp at an input wave of 1.25+$sin(2\pi \cdot 200\text{kHz} \cdot t)$ V. Note we found the *1-dB bandwidth* of the Op-Amp is more sensitive and correlates considerably better to the output signal fidelity of the ET-PA than the more conventional 3-dB bandwidth. As shown in Fig. 5(A), (C) and (E), the output current of the switching stage (I_{SW}) has large ripples on the waveforms, which need to be suppressed or cancelled by the output current of the linear stage (I_{lin}) to reproduce an accurate load current waveform (I_{load}). When the 1-dB bandwidth of the Op-Amp is set as 0.2 MHz, the output voltage (V_{out}) of the SM exhibits not only the switching ripples but also with some attenuation (Fig. 5(B)). When the 1-dB bandwidth is increased to 2 MHz, the V_{out} follows the input voltage (V_{in}) without attenuation, but the switching ripples still cannot be effectively suppressed (Fig. 5(D)). Furthermore, once the 1-dB bandwidth is increased to 8 MHz, the V_{out} follows the V_{in} with high fidelity and negligible ripples (Fig. 5(F)).

To further demonstrate the importance of having a wideband linear stage in the SM to meet the stringent linearity specs, the entire ET-PA using the monolithic SiGe PA is simulated with the RF/analog/digital co-simulation bench in ADS. The behavior model is used for the Op-Amp, while the realistic SPICE models are used for the PA and the other blocks of the SM. The inductor (L) of the buck converter is chosen around 40 µH here. The effect of the value of L on the SM design will be discussed in the next section. The simulated output error-vector-magnitude (EVM) of the ET-PA against different 1-dB bandwidths of the Op-Amp are plotted in Fig. 6 for the WiMAX 64QAM 8.75

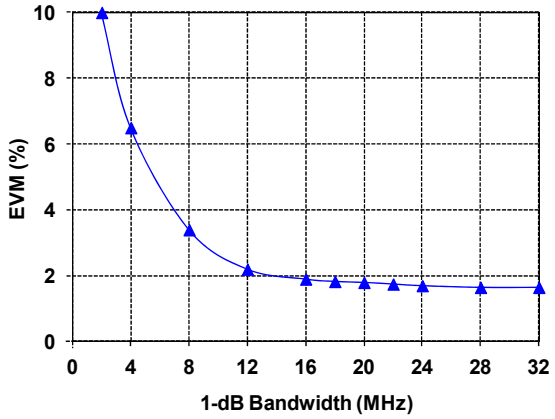


Fig. 6. Simulated output EVM of the ET-PA vs. 1-dB bandwidth of the Op-Amp for the WiMAX 64QAM 8.75 MHz signal. The behavior model is used for the Op-Amp, while realistic SPICE models are used for the SiGe PA and other blocks of the SM, P_{in} = 6 dBm, P_{out} = 16 dBm.

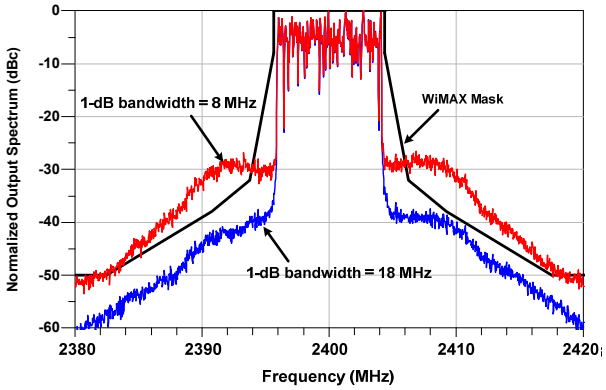


Fig. 7. Simulated output spectra of the ET-PA using different bandwidths of Op-Amp for the WiMAX 64QAM 8.75 MHz signal. The behavior model is used for the Op-Amp, while realistic SPICE models are used for the SiGe PA and other blocks of the SM. P_{in} = 6 dBm, P_{out} = 16 dBm.

MHz. As shown in Fig. 6, the EVM values of the ET-PA decrease as the *1-dB bandwidth* of the Op-Amp increases, and become saturated to ~1.8% after the 1-dB bandwidth of the Op-Amp becomes larger than 18 MHz. Fig. 7 shows the simulated output spectra of the ET-PA with different bandwidths of the Op-Amp. There is a large improvement on the adjacent channel power ratio (ACPR) when the 1-dB bandwidth of the Op-Amp increases from 8 MHz to 18 MHz, enabling the output spectrum passing the stringent WiMAX spectral mask specs. As indicated by Figs. 6-7, the required bandwidth of the SM needs to be at least 2x of the original signal bandwidth, while sufficiently suppressing the switching ripples.

2) *Switching Frequency of the Supply Modulator*

The average switching frequency of the SM shown in Fig. 4 is well analyzed in [7] and can be expressed as [6]:

$$f_{SW} = \frac{R_{sense}}{L} \cdot \frac{V_{dc}}{2h}\left(1 - \frac{V_{rms}^2}{V_{DD} \cdot V_{dc}}\right), \quad (3)$$

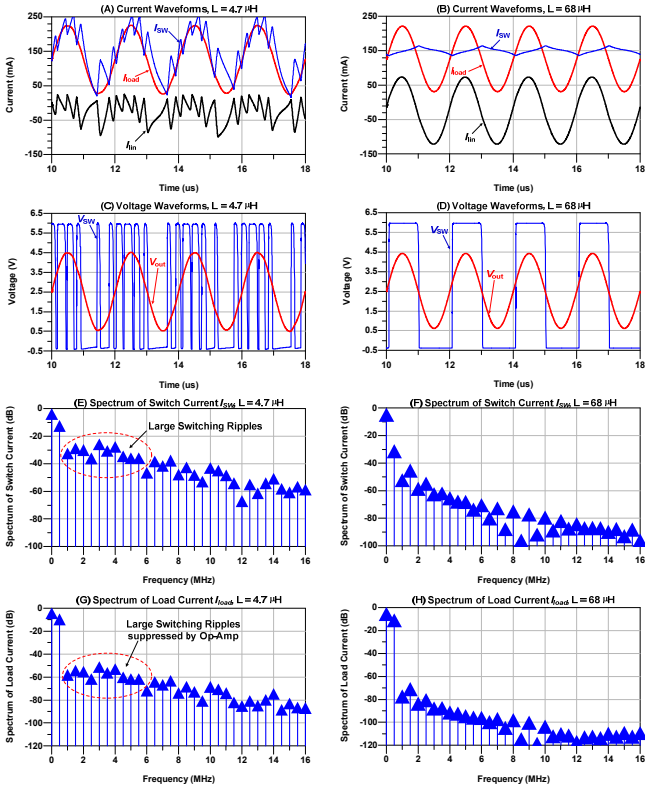


Fig. 8. SPICE simulated (A, B) current waveforms, (C, D) voltage waveforms, (E, F) spectra of the switching current I_{SW}, and (G, H) spectra of the output load current I_{load} of the SM using two different values of L (4.7 µH vs. 68 µH). The realistic SPICE models are used for the supply modulator. R_{load} = 22 Ω, the input voltage = 1.25+$sin(2\pi \cdot 500kHz \cdot t)$ V.

where V_{dc} and V_{rms} are the average and root-mean-square voltages of the output envelope signal, respectively; h is the hysteresis voltage of the comparator. In this design, the comparator LMV7219 has a pre-determined internal hysteresis h of 7-10 mV according to the data sheet and the SPICE simulations. Therefore, according to (3) the average switching frequency can now be mainly controlled by the value of L. The drawback of using a rather small L is that it usually generates more switching ripples at high frequencies, making the design of the linear stage more challenging [17].

Fig. 8 shows the SPICE simulated waveforms and spectra of the SM designed using two different values of L with an input waveform of 1.25+ $sin(2\pi \cdot 500kHz \cdot t)$ V. This time, the realistic SPICE models are used for all blocks of the SM simulations. The switching current I_{sw} supplies both DC and AC components of the load current (I_{load}) by using an L of 4.7 µH; a higher switching frequency and large switching ripples on the waveform of I_{sw} can be observed from Fig. 8(A). Such large switching ripples need to be largely suppressed or cancelled by the output current of the linear Op-Amp (I_{lin}), which can be clearly shown by the spectra of I_{sw} and I_{load} in Figs. 8(E) and (G). On the other hand, for the case of L= 68 µH, I_{sw} supplies only the DC component of I_{load}, while the AC component is taken care of by the linear Op-Amp, as shown in Fig. 8(B). Also, the spectra of I_{sw} and I_{load} for the case of L= 68 µH have smaller harmonics than those using L= 4.7 µH. These simulation

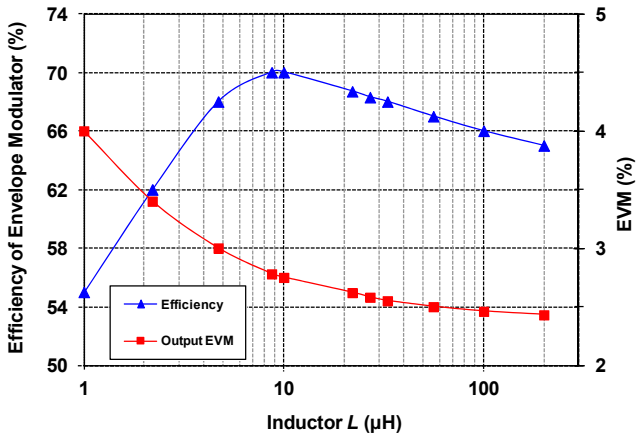


Fig. 9. Simulated efficiency of the SM and output EVM of the ET-PA using different values of L for WiMAX 64QAM 8.75MHz signal. Realistic SPICE models were used for the PA and SM. V_{DD}= 4.2 V, P_{out} = 17 dBm.

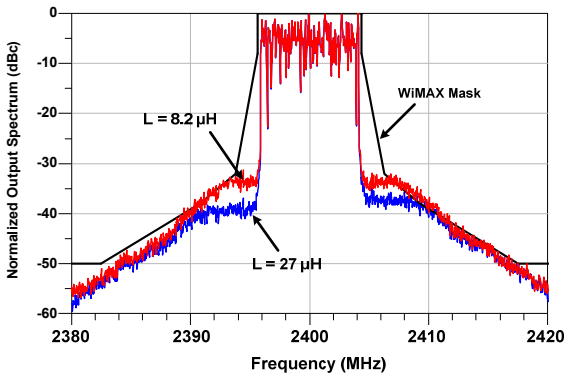


Fig. 10. Simulated output spectra of the ET-PA using different values of L for the WiMAX 64QAM 8.75 MHz signal. Realistic SPICE models were used for the SiGe PA and the EM. V_{DD}= 4.2 V, P_{out} = 17 dBm.

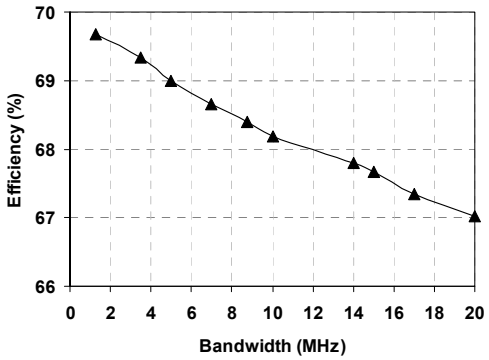


Fig. 11. Measured SM efficiency for different bandwidths of WiMAX 64QAM signals; V_{DD}= 4.2V, R_{load}= 22Ω, average output voltage = 2.3V

results indicate that the optimal L should be selected for not purely the highest efficiency, but also for the sufficient ripple suppression based on the limited bandwidth of the Op-Amp.

Fig. 9 shows the SPICE simulated efficiency of the SM and the EVM of the ET-PA using the SiGe PA (see Fig. 1). The realistic SPICE models are used for the SiGe PA and all blocks of the SM. From the pure view point of efficiency, the optimal L is 8.2 µH for the best efficiency. Smaller L results

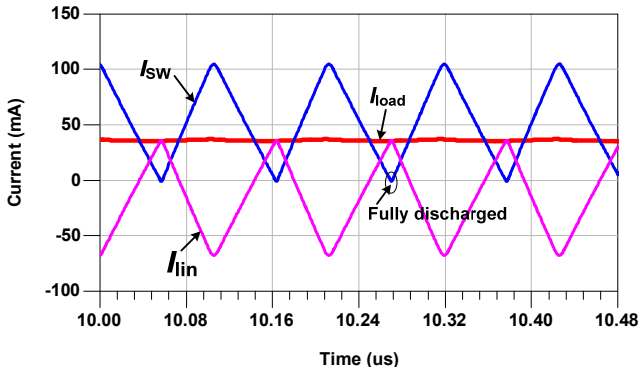


Fig. 12. SPICE simulated current waveforms of the discrete hybrid switching SM at the boundary of the discontinuous mode (L = 1.2 µH)

in higher switching frequency and significant switching loss. In addition, the EVM degrades with smaller L due to higher switching ripples. On the other hand, larger L makes the buck converter only able to supply the DC component of the load current, and in such a case the lower efficiency Op-Amp has to deliver the remaining AC contents (as illustrated in Fig. 8 (B)), leading to lower SM efficiency and thus lower overall efficiency of the ET-PA. A rather large L also causes high parasitic resistance to decrease its efficiency. Fig. 10 shows the SPICE simulated output spectra of the ET-PA using L of 8.2 µH and L of 27 µH, respectively. When the larger L (27 µH) is chosen, the ACPR is 4-6 dB better at the offset of 5-8 MHz from the center frequency with only ~2% lower efficiency (see Fig. 9).

Figs. 9-10 indicate that a small efficiency improvement may not be worthwhile if the linearity of the overall ET-PA has to be sacrificed. Therefore, the L of 27 µH is chosen in the design for our SM to achieve the best trade-off between efficiency and linearity. Fig. 11 shows the measured SM efficiency for different bandwidths of the WiMAX 64QAM signals. The SM efficiency only reduces by 2.5% when the signal bandwidth is increased from 1.5 MHz to 20 MHz.

When the output current is low, the inductor may be completely discharged at the "OFF" state of the buck converter before the switcher is turned on again, which is often called as the "discontinuous mode" for DC-DC converter design [25]. Therefore, another concern in the selection of the inductor value is to ensure the buck converter does not go into the discontinuous mode operation [25]. The boundary of the discontinuous mode occurs at where the output DC current (I_o) equals to one half of the peak-to-peak inductor ramp current ΔI (i.e., $0.5\Delta I = I_o$). For the stand-alone buck converter controlled by the conventional pulse-width modulation (PWM) scheme, the minimal L should be determined to avoid the discontinuous mode at the minimum DC output current ($I_{o,min}$) as [25].

$$L_{min} = \frac{(V_{DD}-V_{out,DC})\cdot D}{\Delta I \cdot f_{SW}} = \frac{(V_{DD}-V_{out,DC})\cdot V_{out,DC}}{2\Delta I_{o,min}\cdot f_{SW}\cdot V_{DD}}, \quad (4)$$

where $V_{out,DC}$ is the output DC voltage, D is the duty cycle, and f_{SW} is the switching frequency determined by the PWM

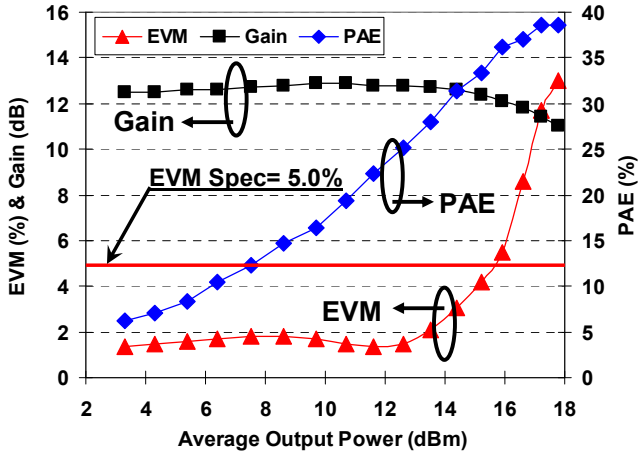


Fig. 13. Measured EVM, gain and PAE vs. average P_{out} of the stand-alone SiGe PA at V_{CC} of 3.6 V for WiMAX 64QAM 8.75MHz signal at 2.3 GHz

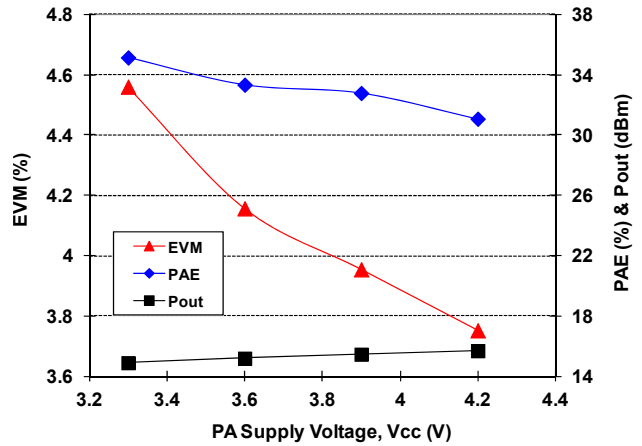


Fig. 14. Measured EVM, PAE and P_{out} vs. V_{CC} of the stand-alone PA for the WiMAX 64QAM 8.75 MHz signal at 2.3 GHz. P_{in} = 3 dBm

control scheme. For example, if the SM were to be implemented using the conventional PWM control scheme for this ET-PA, one could obtain the $V_{out,DC}$ = 2.3 V, $I_{o,min}$ = 33 mA (i.e., at R_{load} = 70 Ω presented by the SiGe PA as shown in Fig. 2). Therefore, the minimal inductor value calculated based on (4) would be ~16 µH for a PWM-controlled buck-converter, assuming f_{SW} =1 MHz.

For the hybrid switching SM in this work, the peak-to-peak inductor ramp current ΔI is limited under $2h/R_{sense}$, which is not related with the inductor value [7]. This is because once the switching current I_{SW} is h/R_{sense} lower than the load current I_{load}, the hysteresis comparator will immediately sense the current difference and turn on the switcher again, assuming the switcher can response fast enough [7]. In the practical SM design, however, the switcher is not ideal due to its intrinsic gate capacitance and resistance, therefore it may not respond fast enough with a high switching frequency, and this frequency is directly determined by the inductor value. In addition, the hysteresis window h increases with higher input slew rate [26]. The SPICE simulations show that h is ~ 7 mV with the input voltage ramp below 0.2 V/µs, but increases to ~ 43 mV with

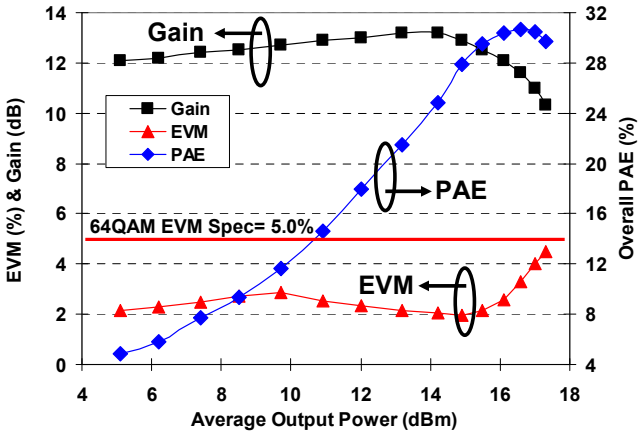


Fig. 15. Measured EVM, gain and overall PAE vs. P_{out} of the ET-PA system for the WiMAX 64QAM 8.75 MHz signal at 2.3 GHz; V_{DD} = 4.2 V

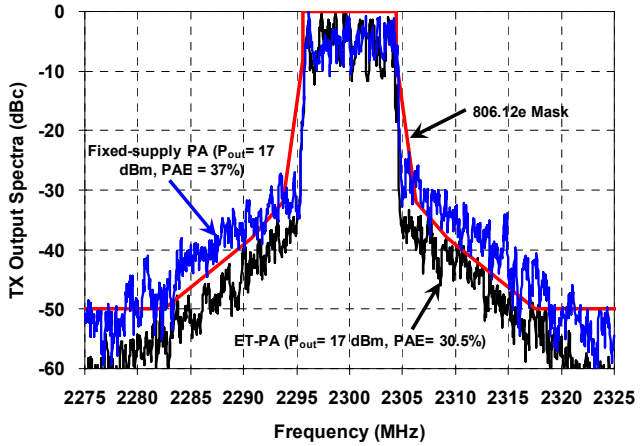


Fig. 16. Measured output spectra of the ET-PA and fixed-supply PA at P_{out} of 17 dBm for WiMAX 64QAM 8.75 MHz at 2.3 GHz

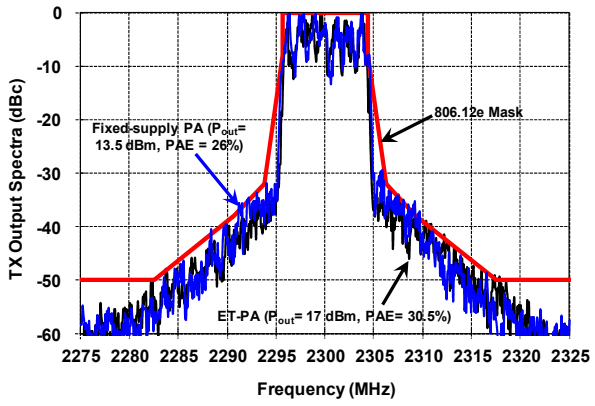


Fig. 17. Measured output spectra of the ET-PA (P_{out}=17 dBm) and fixed-supply PA (P_{out}=13.5 dBm) for WiMAX 64QAM 8.75 MHz at 2.3 GHz

the input voltage ramp of 4 V/µs. According to the simulation, the minimal L is 1.2 µH to avoid the discontinuous mode operation for our SM. Fig. 12 shows the SPICE simulated current waveforms of the hybrid SM at the boundary of the discontinuous mode.

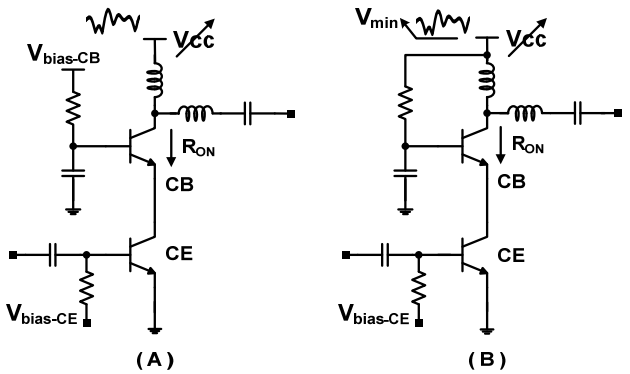


Fig. 18. Simplified schematics: (A) the conventional constant biased cascode SiGe PA; (B) the proposed self-biased cascode SiGe PA

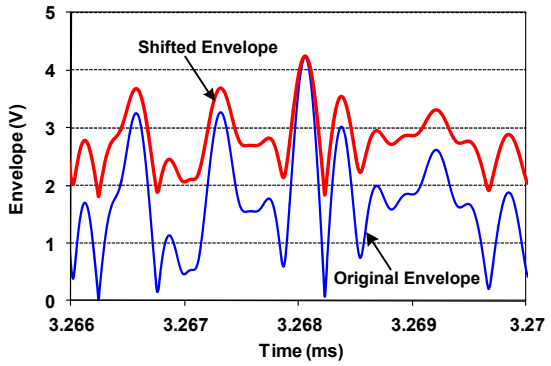


Fig. 19. Simulated envelope waveforms of the WiMAX 64QAM 5 MHz signal before and after using the proposed envelope shifting method

III. COMPARISON OF ET-PA AND STAND-ALONE PA

A. *Efficiency and Linearity of the Stand-Alone PA*

First, the stand-alone SiGe PA with fixed-supply voltage is tested to serve as a reference for the comparison with the ET-PA. Fig. 13 shows the measured gain, PAE and EVM vs. P_{out} for the WiMAX 64QAM 8.75 MHz signal (PAR = 10.5 dB) at 2.3 GHz. The PAE of the SiGe PA reached 39% at the P_{out} of 17.8 dBm, but with a rather high output EVM of 11.7% (the EVM spec of WiMAX 64QAM is 5.0% or -26 dB). At P_{out} of 16 dBm, the stand-alone PA already violates the lenient EVM spec. Increasing V_{CC} could reduce the EVM as shown in Fig. 14, but at the cost of lower efficiency.

B. *Efficiency and Linearity of the ET-PA*

The discrete SM discussed earlier is used to modulate the supply voltage (V_{CC}) of the PA to form an ET-PA. No DPD is used in the measurement. The ET-PA operates at V_{DD} of 4.2 V. Fig. 15 shows the measured EVM, gain and overall PAE of the ET-PA. The overall PAE (or the composite PAE) of the ET-PA includes the power consumption of the SM. The overall PAE of the ET-PA is 30.5% at P_{out} of 17 dBm with an EVM of 4.4%. Fig. 16 shows the output spectra of the ET-PA and the stand-alone PA with a fixed supply. At P_{out} of 17 dBm, the ET-PA passes the stringent WiMAX 64QAM mask, while the fixed-supply PA fails the spectral

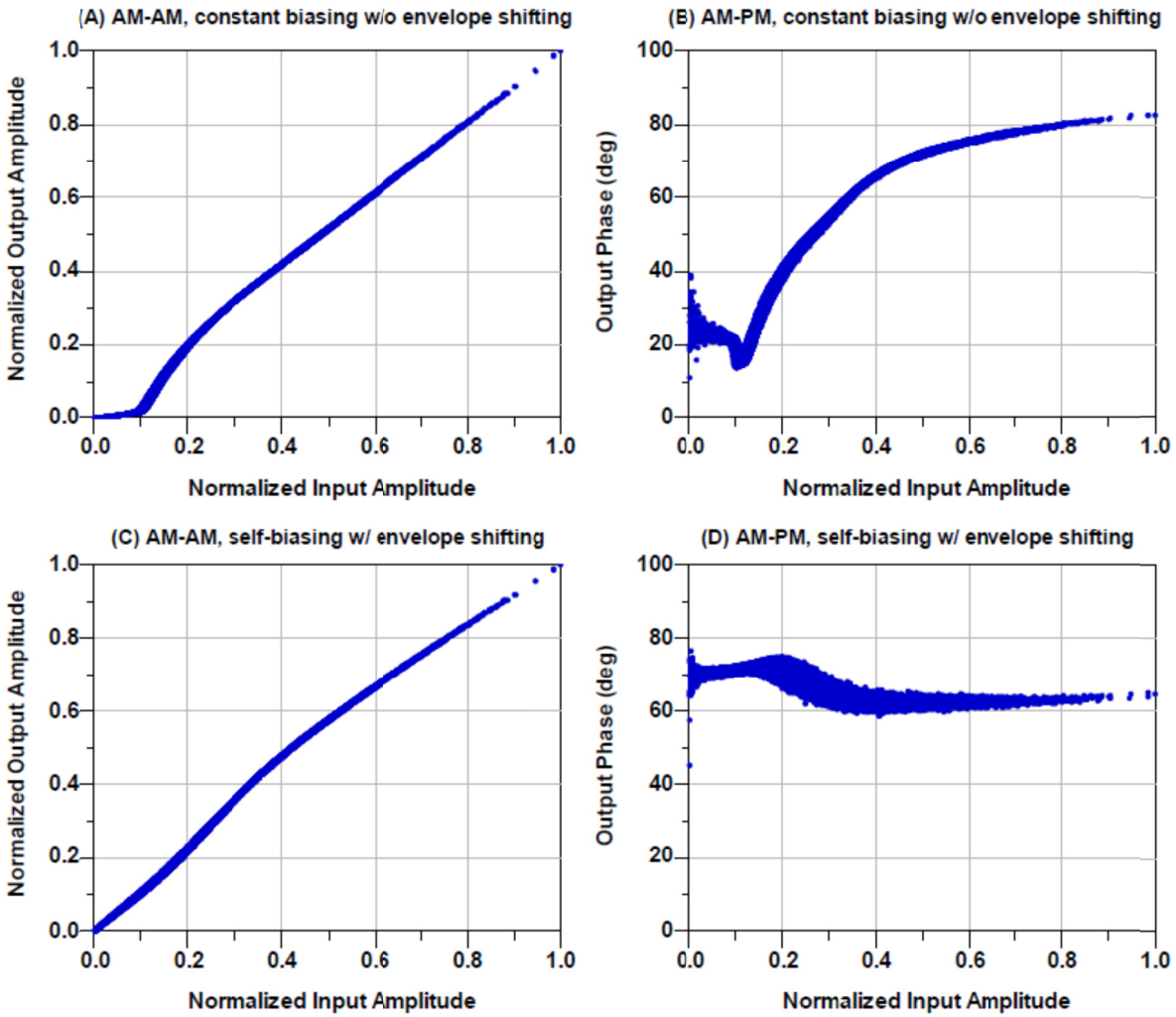


Fig. 20. Simulated AM-AM and AM-PM characteristics of the cascode ET-PA for the WiMAX 64QAM 5 MHz signal: (A, B) conventional constant biased cascode PA without the envelope shifting, (C, D) proposed self-biased cascode PA with the envelope shifting

mask badly. Note that the ET-PA operates at its P_{2dB} point at P_{out} of 17 dBm. Fig. 17 shows that the maximum linear P_{out} of the fixed-supply PA is only ~13.5 dBm in order to pass the WiMAX spectral mask, leading to a PAE of only ~26%.

IV. Fully Monolithic BiCMOS ET-PA

The ET-PA using a SiGe PA and a discrete SM has achieved the promising performances, which prompts us to integrate the SM into the PA to form a low cost solution for handset applications. In this section, a monolithic ET-PA in a 0.35 µm SiGe BiCMOS technology will be presented.

A. Differential Cascode SiGe PA Design

To improve the reliability for Si-based PAs, the cascode topology was used for our PA to relieve the voltage stress on the power devices. In addition, the differential structure is used to reduce the grounding parasitic inductance for higher gain and P_{out}. Although this differential cascode SiGe PA has been presented in [3], here we still highlight its self-biasing structure and our proposed envelope shaping function (i.e., envelope shifting in [3]), since both affect the design of the CMOS SM discussed in the next section.

Fig. 18 shows the simplified schematics of the constant biased cascode PA and the proposed self-biased cascode PA. The envelope waveforms of the WiMAX 64QAM 5 MHz signal before and after the envelope shifting are shown in Fig. 19. Fig. 20 shows the simulated AM-AM and AM-PM characteristics of the ET-PA using the conventional constant biased cascode PA and the proposed self-biased cascode PA with the envelope shifting. In the simulation, the SPICE models provided by the design kit were used for the cascode PA. To only focus on the PA structure and envelope shaping function, an ideal SM (i.e., a gain block without any distortion) is used in this particular simulation to eliminate any distortion from the SM. As shown in Fig. 20, the AM-AM characteristic of the conventional cascode PA under the ET operation has a large distortion at the low instantaneous input amplitude due to the knee effect [10], [27], and its AM-PM characteristic has a large phase difference of ~65°. Both

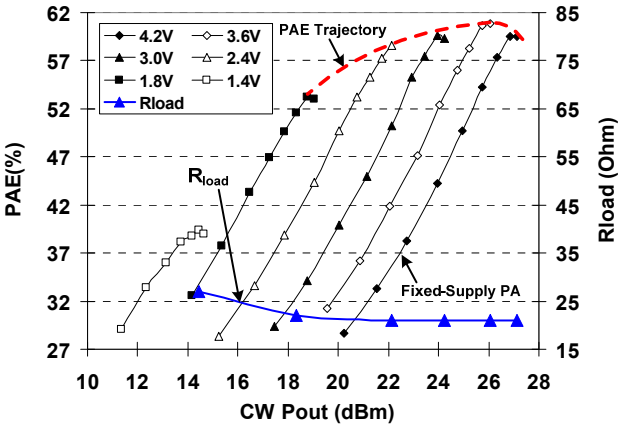


Fig. 21. The PAE vs. P_{out} of the self-biased cascode SiGe PA (measured at different V_{CC} at the CW mode) and the R_{load} seen by the SM

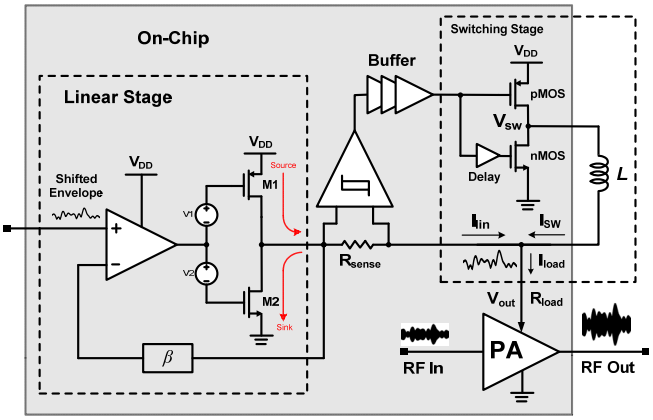


Fig. 22. Simplified block diagram of the CMOS SM integrated with the differential cascode SiGe PA on the same die

AM-AM and AM-PM distortions can be diminished by using the proposed self-biased PA with the envelope shifting.

Following the same procedure as Section II, Fig. 21 shows the measured PAE against P_{out} of the self-biased cascode PA at various V_{CC} in the CW mode. As the V_{CC} is modulated from 1.8 V to 4.2 V, the PAE of the ET-PA varies between 53% and 61%. The R_{load} presented by the self-biased cascode PA stays constantly at ~20 Ω, when V_{CC} is swept from 1.8 V to 4.2 V. The value of R_{load} is important for the analysis of the power losses for the SM.

B. Integrated CMOS Supply Modulator

Fig. 22 shows the simplified block diagram of the CMOS SM integrated with the cascode SiGe PA. The SM was designed in the TSMC 0.35 µm BiCMOS process, but no bipolar devices were used. The SM utilizes the same hybrid switching structure as described in Section II. Although the same SM was presented in [3], [4], its power dissipation has not been discussed in details. In this paper, with our special envelope shaping function and cascode self-biasing PA, we will discuss how our SM can achieve high efficiency with a relatively low average switching frequency.

1) Linear Stage of the Supply Modulator

The linear stage uses a folded cascode amplifier with gain-boosting to meet the slew-rate requirement of the LTE and WiMAX envelope signals [3]. The output stage of the Op-Amp has a common source structure biased at the class-AB mode for a good compromise between distortion and quiescent power dissipation [3]. The efficiency of the class-AB output stage is the key for achieving high efficiency of the linear stage. Since all the DC current is supplied by the switching stage, the loss of the class-AB stage is almost zero at the DC level of the output signal. When the output transistors (M1 and M2) begin to source or sink current, the voltage across them will cause power loss, expressed as:

$$P_{pMOS} = (I_{load} - I_{SW}) \cdot (V_{DD} - V_{out}) \quad (5)$$

$$P_{nMOS} = (I_{SW} - I_{load}) \cdot V_{out} \quad (6)$$

Fig. 23 (A-D) shows the current and voltage waveforms of the supply modulator with and without the envelope shifting for the LTE 16QAM 5 MHz signal. When the load current (I_{load}) is higher than I_{SW}, the pMOS device (M1) of the class-AB stage sources the current; and when I_{load} is lower than I_{SW}, the nMOS device (M2) of the class-AB stage sinks the current. The instantaneous power losses from M1 and M2 against the output voltage (V_{out}) can be clearly seen from Fig. 23 (E) and (F). The average power loss from M1 is reduced from 38 mW to 17 mW by using the envelope shifting technique, while the average power losses from M2 are very close in both cases. According to (5), the envelope shifting mainly reduces the power loss from M1 for the class-AB stage, because it pushes V_{out} closer to V_{DD} (i.e., V_{DD}–V_{out} becomes smaller as shown in Fig. 23 (B)). On the other hand, even though V_{out} becomes higher with the envelope shifting method, the linear stage sinks less current than the case without envelope shifting (see the comparison between Fig. 23 (C) and (D)). This helps to maintain the power loss from M2 roughly the same as the case without using envelope shifting.

2) Switching Stage of the Supply modulator

In this cascode ET-PA system, the envelope shifting method reduces the AC magnitude of the envelope signal, while raising its DC content (see Fig. 19). Therefore, the switching stage supplies more current to the load than the case without envelope shifting, and its efficiency becomes more dominant to the overall efficiency of the supply modulator. It is well-known that there are at least two main mechanisms of power loss in the switching stage: (1) conduction loss; and (2) switching loss. The MOSFET switchers need to be sized for the minimal total power loss (i.e., the conduction loss plus the switching loss). The conduction loss is caused by the on-resistance of the switching FETs when they are conducting, expressed as [11]

$$P_{cond\_loss} = D \cdot \overline{I_{SW}}^2 \cdot R_{on,p} + (1 - D) \cdot \overline{I_{SW}}^2 \cdot R_{on,n}, \quad (7)$$

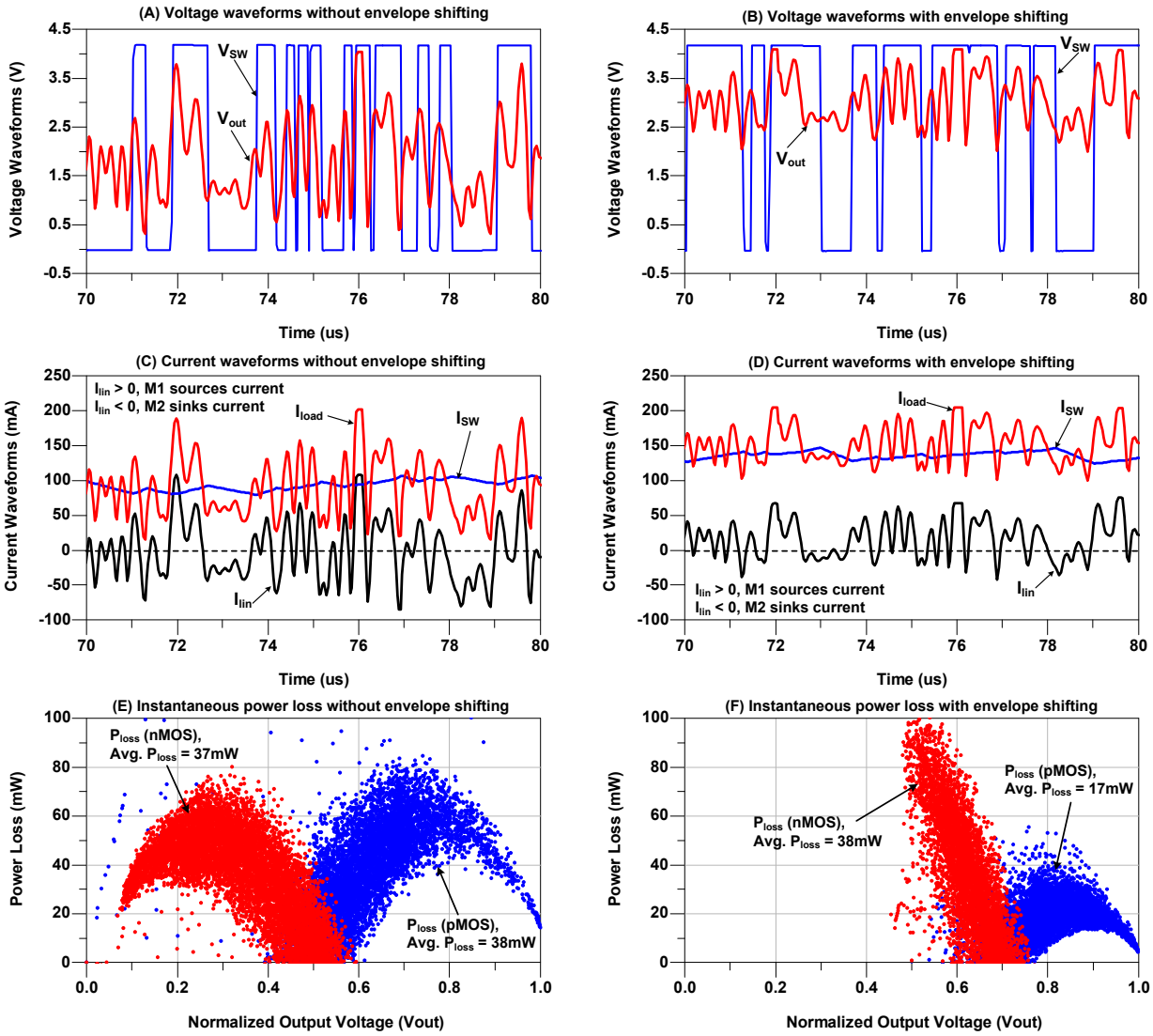


Fig. 23. SPICE simulated waveforms and power losses of the CMOS SM for the LTE 16QAM 5 MHz signal before and after using the envelope shifting method: (A, B) voltage waveforms, (C, D) current waveforms, and (E, F) instantaneous power losses from the class-AB output stage of the Op-Amp. R_{load} = 20 Ω

where D is the average duty cycle of the switching pulses, I_{SW} is the average output current of the buck converter, $R_{on,p}$ is the on-resistance of the pMOS switcher, and $R_{on,n}$ is the on-resistance of the nMOS switcher. The conduction loss is not dependent on the switching frequency and is inversely proportional to the device width. In the design of this integrated CMOS supply modulator, the sizes of pMOS and nMOS of the buck converter are chosen as 20 mm × 0.4 µm and 7 mm × 0.4 µm, respectively. According to the SPICE simulation, $R_{on,p}$ and $R_{on,n}$ are ~0.26 Ω and ~0.27 Ω, respectively. The I_{SW} is 140 mA when driving the R_{load} of 20 Ω presented by the PA. Therefore, the conduction loss of the buck converter is ~10.5 mW according to (7).

The switching loss is caused from the simultaneous current and voltage overlap during the device on/off time (i.e., called crossover loss in [11]), as well as the loss due to its input and output capacitances during switching [7], [9], [11]. The crossover loss can be minimized by adding a delay at the gate of the nMOS as the work in [9] (see Fig. 22). Once the crossover loss is minimized, the switching loss can be expressed as [11]:

$$P_{sw\_loss} = \left(C_d V_{dmax}^2 + C_g V_g^2\right) \cdot f_{SW}, \qquad (8)$$

where C_d is the total drain capacitance, C_g is the total gate capacitance, f_{SW} is the average switching frequency, V_{dmax} is the high level voltage at the drain (i.e., 4.2 V), and V_g is the turn-on voltage of the MOSFET switchers (i.e., 4.2 V). The switching loss is related to both the device width and the switching frequency. A smaller device width can have smaller gate and drain capacitances, but causing higher conduction loss due to its higher on-resistance. For the selected MOSFET switchers, the total drain capacitance and

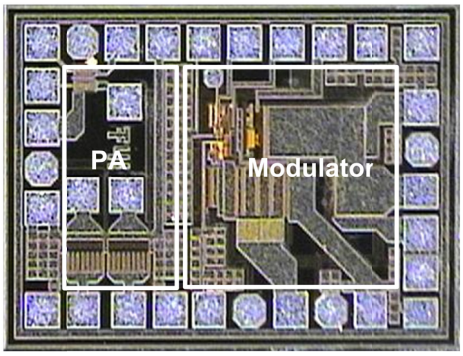


Fig. 24. Chip micrograph of the fully monolithic BiCMOS ET-PA

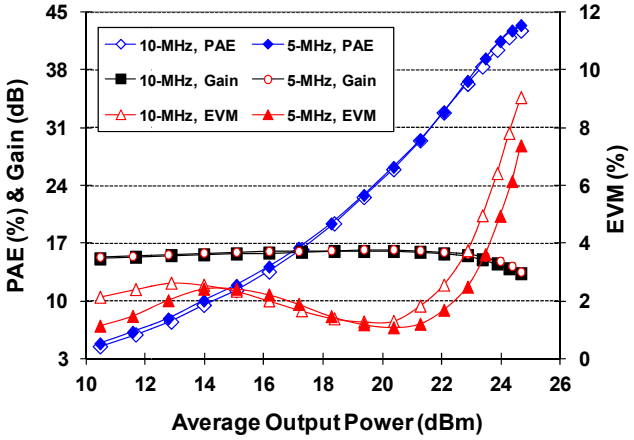


Fig. 25. Measured gain, EVM and overall PAE of the ET-PA for the LTE 16QAM 5 MHz and 10 MHz signals, V_{DD} = 4.2 V

gate capacitance are ~51 pF and ~52 pF obtained from the SPICE simulations, respectively. Additionally, according to the SPICE simulation, the average switching frequency is 0.9 MHz with an inductor of 56 µH. Therefore, the calculated conduction loss is ~1.6 mW from (8). By comparing the values of conduction loss and switching loss, the conduction loss is dominant to the total power loss of the switching stage with a relatively low switching frequency. It is worth noting that the switching frequency of our SM is relatively lower than those of other works [9]-[11], due to the large inductor purposely chosen for low switching ripples and a relaxed linear stage bandwidth as discussed in Section II.

V. PERFORMANCES OF THE FULL MONOLITHIC ET-PA

In this section, we will verify that our design approach of the cascode ET-PA can achieve a high overall efficiency with a relatively low switching frequency of the SM, thus the spectral mask can be satisfied without any need of DPD techniques. The overall ET-PA at V_{DD} of 4.2 V is evaluated for the LTE 16QAM signals with the PAR of 7 dB. The die picture of our fully monolithic ET-PA is shown in Fig. 24, fabricated in the TSMC 0.35 µm SiGe BiCMOS technology. The total chip size is 1.5 × 1.1 mm^2. No DPD will be used in the measurement.

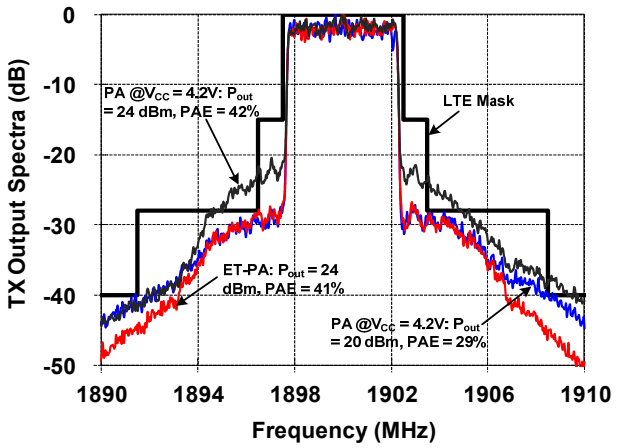


Fig. 26. Measured output spectra of the ET-PA and the stand-alone fixed supply PA for the LTE 16QAM 5 MHz signal

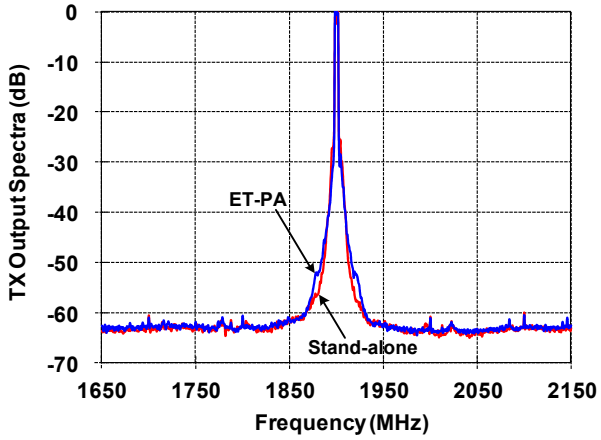


Fig. 27. Measured far-out output spectra of the ET-PA and stand-alone PA for the LTE 16QAM 5 MHz signal, both at P_{out} of 24 dBm

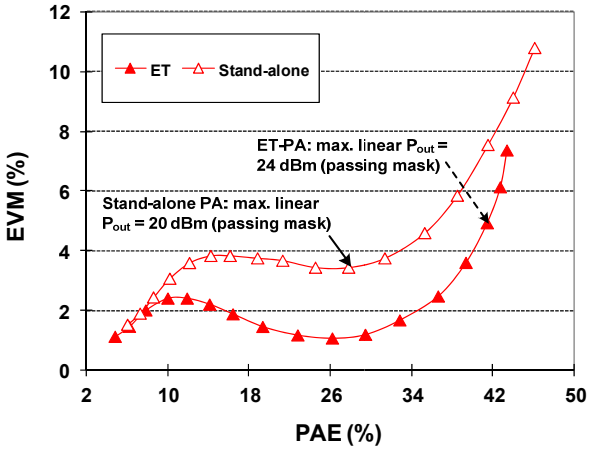


Fig. 28. Efficiency-linearity comparison for the ET-PA and the stand-alone PA for the LTE 16QAM 5 MHz signal

Fig. 25 shows the measured EVM, gain and overall PAE of the ET-PA for both the LTE 16QAM 5 MHz and 10 MHz signals at 1.9 GHz. While keeping the EVM below 5%, the

TABLE I. SUMMARY OF OUR FULLY MONOLITHIC ET-PA AND THE COMPARISON WITH STATE-OF-THE-ART ET/POLAR PAS

	Freq. (GHz)	V_{DD} (V)	P_{out} (dBm)	*Overall PAE	EVM	Modulation	DPD	Technology
[9]	2.4	3.3	20	28%	5%	WLAN 64QAM 20 MHz	Yes	0.18 µm SiGe BiCMOS
[10]	1.88	3.3	29	46%	---	WCDMA 3.84 MHz	No	PA: 2 µm InGaP/GaAs
	1.88	3.3	23.9	34.3%	2.98%	WiMAX 64QAM 5	No	SM: 0.13 µm CMOS
[12]	2.535	6	29	43%	1.9%	LTE 16QAM 20 MHz	Yes	PA: GaAs HBT SM: 0.15 µm CMOS
[22]	2.0	3.3	19.6	22.6%	2.5%	WLAN 64QAM 20 MHz	No	0.13 µm CMOS
[24]	1.92	2.1	15.3	22%	1.5%	WiMAX 64QAM 5 MHz	Yes	0.13 µm SOI-CMOS
[27]	1.85	5	28.9	42.2%	2.69%	LTE 16QAM 10 MHz	No	PA: 2 µm InGaP/GaAs SM: 0.18 µm CMOS
[28]	1.88	3	24.22	38.6%	3.64%	WiBro 16QAM 5 MHz	No	PA: 2 µm InGaP/GaAs SM: 0.13 µm CMOS
[29]	2.535	3.6	25.8	32.3%	3%	LTE 16QAM 10 MHz	No	PA: 2 µm InGaP/GaAs SM: 0.35 µm CMOS
[30]	2.35	3.5	24.7	25.5	4.5	LTE 16QAM 20 MHz	No	0.18 µm SiGe BiCMOS
This work	1.9	4.2	24	41%	4.9%	LTE 16QAM 5 MHz	No	0.35 µm SiGe BiCMOS
			23.4	38%	4.9%	LTE 16QAM 10 MHz		

Note: * Overall PAE includes the efficiency of the SM

maximum linear P_{out} is 24 dBm with an overall PAE of 41% for the LTE 5 MHz signal. In order to achieve the same EVM for the LTE 10 MHz signal, the maximum linear P_{out} needs to be 0.6 dB lower than the case of LTE 5 MHz, resulting in an overall PAE of 38% at P_{out} of 23.4 dBm. When the ET-PA is backed off 0.5 dB from the maximum linear P_{out} levels, better EVMs of 3.6% and 3.7% can be achieved with overall PAEs of 39% and 37% for the LTE 16QAM 5 MHz and 10 MHz signals, respectively.

The output spectra of the ET-PA and the stand-alone fixed supply PA (V_{CC} = 4.2 V) are plotted in Fig. 26. The ET-PA successfully passes the LTE spectral mask at P_{out} of 24 dBm. However, at the same P_{out} of 24 dBm, the stand-alone fixed-supply PA fails the LTE spectral mask, forcing it to be backed off by at least 4 dB. The maximum linear P_{out} of the stand-alone PA is, therefore, only 20 dBm with a PAE of 29%. The linear P_{out} and PAE are thus 4 dB and 12% lower than those achieved by the ET-PA, respectively. Fig. 27 plots the far-out output spectra of the ET-PA and the stand-alone PA measured without any external filtering. Compared with the stand-alone PA, the ET-PA has little spectral regrowth at the offset of 20-30 MHz from the center frequency, but no strong spur appears at the offset frequency above 50 MHz. Such a spurious emission performance was achieved due to the low average switching frequency of the SM. To further demonstrate the advantages of the ET-PA, Fig. 28 compares the measured EVM of the ET-PA and the stand-alone PA against the PAE values. The ET-PA has a better EVM with the same PAE. Table I provides the performance summary of our fully monolithic ET-PA together with the state-of-the-art results.

VI. CONCLUSION

The complete design path towards our fully monolithic BiCMOS ET-PA has been presented. A discrete hybrid switching SM was first implemented to study the effects of its switching frequency and bandwidth limitation to the overall ET performance. The same circuit configuration was used to design our CMOS SM integrated with the self-biased cascode SiGe PA under our proposed envelope shaping function. By analyzing the power losses of the CMOS SM for our special ET operation, we showed that high efficiency can still be achieved with a relatively low average switching frequency, which helped to improve the overall ET linearity. At 1.9 GHz, our BiCMOS ET-PA achieved 24/23.4 dBm with the overall PAE of 41%/38% for the LTE 16QAM 5/10 MHz signals, respectively. The EVM below 5% and the LTE spectral mask were both satisfied without any need of DPD techniques. The literature survey indicates that our design achieved one of the highest efficiency for Si-based ET PAs, approaching those III-V semiconductor PAs for broadband mobile applications.

ACKNOWLEDGMENT

The authors are deeply grateful to the Industrial Technology Research Institute (ITRI), Taiwan, R.O.C. for research funding support and TSMC for IC fabrication. The authors would also like to thank Dr. Keh-Shew Lu, CEO of Diodes Inc. for contributing and setting up the Keh-Shew Lu Regents Endowment Fund at Texas Tech University.

REFERENCES

[1] Y. Li, J. Lopez, and D. Y.C. Lie, "A wideband envelope modulator design for envelope-tracking SiGe power amplifier (ET-PA) for broadband wireless applications," Proc. 10th Int'l Conf. Wireless & Mobile Comm., Jun. 2014, pp. 76-83.

[2] J. Lopez, Y. Li, J.D. Popp, D.Y.C. Lie, C.C. Chuang, K. Chen, S. Wu, T-Y. Ying, and G-K Ma, "Design of highly efficient wideband RF polar transmitter using the envelope-tracking technique," IEEE J. Solid-State Circuits, vol. 44, no. 9, pp. 2276-2294, Sept. 2009.

[3] Y. Li, J. Lopez, P.-H. Wu, W. Hu, R. Wu, and D.Y.C. Lie, "A SiGe envelope-tracking power amplifier with an integrated CMOS envelope modulator for mobile WiMAX/3GPP LTE transmitters," IEEE Trans. Microw. Theory Tech., vol. 59, no. 10, pp. 2525-2536, Oct. 2011.

[4] Y. Li, J. Lopez, C. Schecht, R. Wu, and D. Y.C. Lie, "Design of high efficiency monolithic power amplifier with envelope-tracking and transistor resizing for broadband wireless applications," IEEE J. Solid-State Circuits, vol. 47, no. 9, pp. 2007-2018, Sept. 2012.

[5] D. Y.C. Lie, J. Lopez, J. D. Popp, J. F. Rowland, G. Wang, G. Qin, and Z. Ma, "Highly-efficient monolithic class E SiGe power amplifier design at 900 and 2400MHz," IEEE Trans. Circuits Syst. I – Reg. Papers, vol. 56, no. 7, pp. 1455-1466, Jul. 2009.

[6] Y. Li, J. Lopez, D.Y.C. Lie, K. Chen, S. Wu, T.-Y. Yang, and G.-K. Ma, "Circuits and system design of RF polar transmitters using envelope-tracking and SiGe power amplifier for mobile WiMAX," IEEE Trans. Circuits Syst. I – Reg. Papers, vol. 58, no. 5, pp. 893-901, May 2011.

[7] F. Wang, D.F. Kimball, J.D. Popp, A.H. Yang, D.Y.C. Lie, P.M. Asbeck and L.E. Larson, "An improved power-added efficiency 19-dBm hybird envelope elimination and restoration power amplifier for 802.11g WLAN application," IEEE Trans. Microw. Theory Tech., vol. 54, no. 12, pp. 4086-4099, Dec. 2006.

[8] F. Wang, A. H. Yang, D. F. Kimball, L E. Larson, and P. M. Asbeck, "Design of wide-bandwidth Envelope-Tracking power amplifiers for OFDM applications," IEEE Tran. Microw. Theory Tech., vol. 53, no. 4, pp. 1244-1255, Apr. 2005.

[9] F. Wang, D. Kimball, D. Y.C. Lie, P. Asbeck, and L. E. Larson, "A monolithic high-efficiency 2.4-GHz 20-dBm SiGe BiCMOS envelope-tracking OFDM power amplifier," IEEE J. Solid-State Circuits, vol. 42, no. 6, pp. 1271-1281, Jun. 2007.

[10] J. Choi, D. Kim, D. Kang, and B. Kim, "A polar transmitter with CMOS programmable hysteretic-controlled hybrid switching supply modulator for multistandard applications," IEEE Trans. Microw. Theory Tech., vol. 57, no.7, pp. 1675-1686, Jul. 2009.

[11] M. Kwak, D. F. Kimball, C. D. Presti, A. Scuderi, C. Santagati, J. J. Yan, P. M. Asbeck, and L. E. Larson, "Design of a wideband high-voltage high-efficiency BiCMOS envelope amplifier for micro-base-station RF power amplifier," IEEE Trans. Microw. Theory Tech., vol. 60, pp. 1850-1861, Jun. 2012.

[12] M. Hassan, L. E. Larson, V. W. Leung, D. F. Kimball, and P. M. Asbeck, "A wideband CMOS/GaAs HBT envelope tracking power amplifier for 4G LTE mobile terminal applications," IEEE Trans. Microw. Theory Tech., vol. 60, no. 5, pp. 1321–1330, May 2012.

[13] B. J. Minnis, P. A. Moore, P. N. Whatmough, P. G. Blanken, and M. P. van der Heijden, "System-efficiency analysis of power amplifier supply-tracking regimes in mobile transmitters," IEEE Trans. Circuits Syst. I – Reg. Paper, 56, 1, pp. 268-279, Jan. 2009.

[14] J. Jeong, D. F. Kimball, M. Kwak, C. Hsia, P. Draxler, and P. M. Asbeck, "Wideband envelope tracking power amplifiers with reduced bandwidth power supply waveforms and adaptive digital predistortion techniques," IEEE Tran. Microw. Theory Tech., vol. 53, no. 4, pp. 1244-1255, Apr. 2005.

[15] J. Kitchen, W. Chu, I. Deligoz, S. Kiaei, and B. Bakkaloglu, "Combined linear and Δ-modulated switch-mode PA supply modulator for polar transmitters", IEEE Int. Solid-State Circuits Conf. Tech. Dig., Feb. 2007, pp. 82-83.

[16] F. H. Raab, "Split-band modulator for Kahn-technique transmitters," IEEE MTT-S Int. Microw. Symp. Dig., 2001, pp. 887-890.

[17] R. Shrestha, R. v.d. Zee, A. d. Graauw, and B. Nauta, "A wideband supply modulator for 20 MHz RF bandwidth polar PAs in 65 nm CMOS," IEEE J. Solid-State Circuits, vol. 44, no. 4, pp. 1272-1280, Apr. 2009.

[18] P. Reynaert and M. S.J. Steyaert, "A 1.75-GHz polar modulated CMOS RF power amplifier for GSM-EDGE," IEEE J. Solid-State Circuits, vol. 40, no. 12, pp. 2598-2608, Dec. 2005.

[19] D. K. Su and W. J. McFarland, "An IC for linearizing RF power amplifier using envelope elimination and restoration," IEEE J. Solid-State Circuits, vol. 33, no. 12, pp. 2252-2258, Dec. 1998.

[20] V. Pinon, F. Hasbani, A. Giry, D. Pache, and C. Garnier, "A single-chip WCDMA envelope reconstruction LDMOS PA with 130 MHz switched-mode power supply," IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers, Feb. 2008, pp. 564–565.

[21] M. Helaoui, S. Boumaiza, A. Chazel, and F. M. Ghannouchi, "On the RF/DSP design for efficiency of OFDM transmitters", IEEE Trans. Microw. Theory Tech., 53, 7, pp. 2355-2361, 2005.

[22] J. S. Walling, S. S. Taylor, and D. J. Allstot, "A class-G supply modulator and class-E PA in 130nm CMOS," IEEE J. Solid-State Circuits, vol. 44, no. 9, pp. 2339-2347, Sept. 2009.

[23] A. Kavousian, D. K. Su, M. Hekmat, A. Shirvani, and B. A. Wooley, "A digitally modulated polar CMOS power amplifier with a 20-MHz channel bandwidth," IEEE J. Solid-State Circuits, vol. 43, no. 10, pp. 2251-2258, Oct. 2008.

[24] C. D. Presti, F. Carrara, A. Scuderi, P. M. Asbeck, and G. Palmisano, "A 25 dBm digitally modulated CMOS power amplifier for WCDMA/EDGE/OFDM with adaptive digital predistortion and efficient power control," IEEE J. Solid-State Circuits, vol. 44, no. 7, pp. 1883-1896, Jul. 2009.

[25] A. I. Pressman, Switching Power Supply Design, Second Edition, New York, NY: McGraw Hill, 1998.

[26] R. Gregorian, Introduction to CMOS Op-Amps and Comparators, New York, NY: Wiley, 1999.

[27] D. Kim, D. Kang, J. Choi, J. Kim, Y. Cho, and B. Kim, "Optimization for envelope shaped operation of envelope tracking power amplifier," IEEE Trans. Microw. Theory Tech., vol. 59, no. 7, pp. 1787-1795, Jul. 2011.

[28] J. Choi, D. Kang, D. Kim, and K. Kim, "Optimized envelope tracking operation of Doherty power amplifier for high efficiency over an extended dynamic range," IEEE Trans. Microw. Theory Tech., vol. 57, no. 6, pp. 1508-1515, Jun. 2009.

[29] J. Choi, D. Kim, D. Kang and K. Kim, "A new power management IC architecture for envelope tracking power amplifier," IEEE Trans. Microw. Theory Tech., vol. 59, no. 7, pp. 1796-1802, Jul. 2011.

[30] M.-L. Lee, C.-Y. Liou, W.-T. Tsai, C.-Y. Lou, H.-L. Hsu, and S.-G. Mao, "Fully monolithic BiCMOS reconfigurable power amplifier for multi-mode and multi-band applications," IEEE Trans. Microw. Theory Tech., vol. 63, no. 2, pp. 614–624, Feb. 2015.

Performance Comparison of Data Serialization Schemes for ETSI ITS Car-to-X Communication Systems

Sebastian Bittl, Arturo A. Gonzalez, Michael Spähn, and Wolf A. Heidrich
Fraunhofer ESK
Munich, Germany
Email: {sebastian.bittl, arturo.gonzalez, michael.spaehn, wolf.heidrich}@esk.fraunhofer.de

***Abstract*—Wireless Car-to-X communication is about to enter the mass market in upcoming years. The non-licensed, but reserved bandwidth for such communications is relatively narrow in terms of the few usable channels and the expected number of communicating entities. Among other communication aspects, data serialization schemes that allow compact encoding as well as fast encoding and decoding are required. Compact representation of data helps in keeping a minimal bandwidth overhead in the wireless channel. Moreover, since delay is an important performance parameter in Car-to-X communication, the processing time of the serialization/deserialization schemes shall be kept to a minimum. So far, no detailed analysis of different data serialization schemes for Car-to-X communication regarding important properties such as runtime, memory consumption and encoded output length has been published. In this work, we provide a performance comparison analysis between the standardized ASN.1 , the binary representations, Google Protocol Buffers and Efficient XML Interchange format (EXI), all of them as alternative strategies for data serialization. Standardized data content for CAM, DENM and the security envelope are used in the conducted study. We conclude that ASN.1 encoding on the facility layer shows the best performance, outperforming Google Protocol Buffers and EXI. However, for the case of encoding the security envelope, ASN.1 is outperformed by a binary encoding scheme in most cases, while EXI encoding outperforms all other schemes. This implies that standardization efforts for the security envelope should reconsider the recent shift from binary encoding towards usage of ASN.1. Instead, the Efficient XML Interchange format should be considered for this purpose.**

Keywords-ETSI ITS; data serialization; ASN.1; Google Protocol Buffers; Efficient XML Interchange format.

I. INTRODUCTION

Car-to-X (C2X) communication systems, often called Vehicular Ad-Hoc Networks (VANETs), are gaining increased attention in the awake of their upcoming deployment. Considering the high number of vehicles that are to be interconnected and the narrow radio spectrum being a limited resource for communication, efficient data representation is an important aspect, which has not received much attention prior to our work in [1].

However, the kind of data representation used for sending content over the wireless channel significantly influences the overall network performance. One of the key impacts is the one on the availability of the channel, since a longer packet implies a longer usage of the spectrum at a given transmission rate. Since the ITS-G5 and WAVE systems both use a carrier sense collision avoidance mechanism (CSMA-CA) for multi-user channel access, longer packets imply longer transmission times and hence a higher channel busy ratio. This means that nodes near a transmitter, which are willing to transmit as well, must wait longer until the channel is free to use. Moreover, the wireless channel has a limited capacity allowing only a limited number of transmitted bits per unit of time. The lower the number of transmitted messages, the higher the number of users that could use the given frequency spectrum in parallel.

The data serialization schemes also influence processing power requirements in both ETSI Intelligent Transport Systems (ITS) in Europe [2] and Wireless Access in Vehicular Environments (WAVE) in the United States [3].

Like most modern communication systems, C2X communication happens digitally. The latter implies that messages between the involved nodes are represented as a series of bits, i.e., as bit streams, in a platform independent way. As in any software implementation of a communication system, the format of the messages exchanged between two communication end points must be well known by them. That means that nodes should be able to represent messages as bit streams and to interpret them as the original messages as well. The generation of a bit stream from a message is defined as encoding. Hence, we refer to an encoded message as the bit stream representation of such a message. Following the same logic, decoding is defined as the (re-)generation of the original message from its bit stream representation.

Several encoding schemes exist nowadays, and some of them are used extensively in everyday data communications. Depending on the application requirements, one scheme may be suited better than another. The requirements for these encoding schemes range from human readability (e.g., XML [4], JSON [5]) or the space the bit stream takes up in memory (e.g., ASN.1 PER encoding, binary encoding), up to system performance, i.e., encoding/decoding processing delay (e.g., binary encoding, ASN.1 OER encoding).

In the C2X realm, it is significantly relevant to use a bandwidth efficient encoding scheme since C2X communications operate under quite strict bandwidth constraints. As an example, in Europe only one 10 MHz channel is available for safety critical applications [6]. Therefore, an encoding scheme that generates short bit streams out of messages is favored. Moreover, safety related C2X applications have strict end-to-end delay requirements. Therefore, encoding/decoding delays

should be minimal such that their contribution to the end-to-end delay may be considered negligible.

In this work, we focus on the comparison of the performance metrics of three coding schemes, namely Abstract Syntax Notation 1 (ASN.1) encoding rules, Google Protocol Buffers (protobuf) and the Efficient XML Interchange format (EXI). We apply the mentioned data serialization schemes to the two most common C2X message types in ETSI ITS systems, which are Cooperative Awareness Message (CAM) and Decentralized Environmental Notification Message (DENM). Furthermore, four different encoding schemes for the ETSI ITS security envelope are compared, which are binary encoding, ASN.1 encoding rules, protobuf and EXI.

The remaining part of this work is organized as follows. An overview of related work is given in Section II. Afterwards, performance requirements and the applied measurement mechanisms are described in detail in Section III. The target platforms' description is summarized in Section IV. The obtained results of the extensive performance study are described in Section V. Finally, Section VI provides a conclusion about the achieved results.

II. Background

The background of this work regarding platform independent data encoding, especially in the area of ETSI ITS, is provided in this section. Additionally, a comparison to the limited number of other published performance studies in the area of this work is given.

A. Data Encoding Rules

Cooperative Awareness Messages (CAMs) and Decentralized Environmental Notification Messages (DENMs) are the two most important standardized ETSI ITS messages defined in [7] and [8], respectively. A CAM contains basic information about the transmitting vehicle, for example position and speed. Depending on the purpose of the C2X application, received CAMs are processed in specific ways. One example of an application consuming the information contained in CAMs, is the use case of a C2X based collision risk warning. Here, the receiver nodes asses the risk of a collision with the transmitter given the speed, position and heading information of the sender, which is included in the CAM. The CAM generation period is defined in the standard to be between 1 and 10 Hz, being able to adapt depending on several parameters, such as vehicle velocity and channel busy ratio [7].

In contrast to CAMs, DENMs are event-triggered messages. As its name suggest, a DENM contains information about an event in the vicinity of the originating ITS-Station (ITS-S). One example of an event could be a stationary vehicle on the road. In this case, the stationary vehicle generates a DENM and depending on the configuration might broadcast this message periodically afterwards. The contents of the DENM include, among other information, the position and type of the event [8].

According to ETSI ITS standards, the encoding of CAM and DENM is done using ASN.1 UPER (Unaligned Packed Encoding Rules) encoding rules. There are several encoding rules specified for ASN.1, including Basic Encoding Rules (BER), Packed Encoding Rules (PER), Canonical Encoding Rules (CER), Distinguished Encoding Rules (DER), Octet Encoding Rules (OER) and XML Encoding Rules (XER) among many other flavors. Each of them is providing advantages and disadvantages from the point of view of a specific application.

Since PER provides a more compact encoded message than the older BER and its subsets DER and CER, it is often used in systems where bandwidth conservation is important [9]. This might be the reason why the CAM and DENM standards specify that the encoding rules to be used should be Unaligned PER (UPER). In aligned PER fields are aligned to 8-bit octet boundaries by inserting padding bits, whereas in UPER padding bits are never inserted between fields, hence allowing a higher bit stream size reduction.

A widely used alternative to ASN.1 encoding rules are the so called Google Protocol Buffers (protobuf), which are used by Google extensively in their production environment [10], [11]. Therefore, they can be regarded as a stable and reliable mechanism. Protobuf offers a more simplistic approach to the platform independent data encoding, making it easier to manipulate and implement [10]. Additionally, they can be configured to do encoding optimized for either fast processing or small memory footprint. The latter is also a common feature provided by standard ASN.1 implementations. For example, the software provided by OSS Nokalva provides this feature [9]. Therefore, protobuf can be seen a well comparable alternative to ASN.1 for data representation. Hence, the performance study provided in Section V includes usage of protobuf.

Another data representation technology, which is well-known in the automotive domain, is the so called binary Extensible Markup Language (XML) via the Efficient XML Interchange (EXI) standard [12]. EXI is used for the communication between electric vehicles and charging stations during the charging process as defined in the new ISO 15118 standard [13]. EXI is a machine-to-machine protocol, which aims in removing all overhead coming along with XML, to enable human readability such as indentation, whitespace and even byte alignment (like for ASN.1 UPER). In schema less mode, EXI tries to minimize overhead by the use of string tables, thus reduces encoding size very rapidly for repetitive message structures.

Moreover, EXI also allows the usage of XML schemes on the sender and the receiver, which allows to share common knowledge about the structure of the transferred messages in a convenient and flexible way. EXI encoding and decoding happens in a linear way, so that decoding can already start when only a first part of the full message has been received. Thus, EXI can be regarded a proper alternative to ASN.1 UPER and is therefore included in our performance study.

For the ETSI ITS security envelope two different sets of encoding rules have been proposed so far. At first, binary encoding with explicit definition of all data fields was proposed in [14]. Additionally, encoding using ASN.1 UPER was proposed recently in [15]. However, no in detail comparison of these two proposals is available, except of our prior work in [1]. As further reference schemes, protobuf and EXI are also used for the security envelope in the performance study, presented in Section V.

There are important differences to consider when comparing different data representation schemes. Some of the most relevant are,

- how are optional fields within messages handled, i.e., how is a field's presence or absence represented,
- possibility of future backward compatibility when extending a message, i.e., adding of new mandatory or optional data fields,
- byte alignment,
- providing the functionality of data compression, for example variable length representation of integers.

These differences allow many different combinations which could be used and are actually to be found in various practically used mechanisms. An overview of this set of basic properties for the data representation schemes used later in this work is provided in Table I.

TABLE I. OVERVIEW OF BASIC PROPERTIES FOR DIFFERENT SCHEMES OF PLATFORM INDEPENDENT DATA REPRESENTATION.

	binary [14]	ASN.1 UPER	protobuf [16]	EXI
presence of optional fields	not encoded	encoded	encoded	encoded
extendability	no	no	yes	yes
byte alignment	yes	no	yes	no
compression	normally no only one type	yes	yes byte blocks	yes byte blocks

As a comparison example, the standardized binary encoding for the ETSI ITS security envelope uses no explicit encoding of optional fields [14], allowing no possibility of a backward compatible extension. It also uses data compression only for one specific data type (IntX [14]), which is a variable length integer type. This type is almost exclusively used to encode the length of data fields and not for encoding real data content. Protobuf and EXI both use variable length encoding for integers, but the size of an integer still has to be an integer multiple of one byte.

In regard to extendability, protobuf provides a backward compatibility property, which allows to add message structures without any change to the encoding and decoding code. EXI has a similar feature when using schema files for message specification. Here the schema file might be updated in a way that the old message specification is still valid (e.g., by adding an optional element). In this case, no change in the encoding and decoding code is necessary either.

Alternative publicly available data serialization tools for converting arbitrary data into a platform independent binary representation include systems like Apache Avro, Apache Thrift or Message Pack [17]–[19]. These systems are either less mature or deployed to a much smaller extent in professional environments compared to ASN.1 and protobuf (e.g., see [20] for protobuf vs. Thrift). Therefore, they are not studied in detail in this work. Additionally, serialization technologies like XML or JSON, which aim to achieve a human readable and easy to parse data representation at the price of increased encoding length are out of the scope of this work. Such systems are not appropriate to be used in bandwidth constrained communication systems.

B. State of the Art and Contribution of this Work

There are several publications comparing other encoding schemes, such as XML with ASN.1. For example, the authors in [21] compare the performance between binary encoded XML and ASN.1 by running the tests on PC machines. In [22], the authors compare the performance of XML against ASN.1 BER on digitally signed data. They conclude that for applications where high performance is required, ASN.1 BER may be a better choice.

In [23] authors compare the performance of XML, JSON and protobuf in terms of data size and coding speed. The authors conclude that protobuf requires less bytes for the message representation in comparison with XML or JSON. The authors also explore the possibility of compressing XML and JSON messages using gzip [24]. In the latter case, both compressed text formats perform better than protobuf in terms of data size. In terms of speed, the authors show that protobuf performs better than both text schemes.

In [25], authors perform a similar study to the one in [23] and expand it for performance in energy consumption, relevant for a smartphone use case. They also show that gzip-compressed protobuf, a variant not explored in [23], performs better in terms of encoded data size in comparison with compressed XML, but worse than compressed JSON. When the authors measure performance in respect to encoding time, they conclude that for the data set they used protobuf performs better. On the parsing process on the receiver side, i.e., decoding, JSON performs slightly better than the other two schemes.

A performance comparison between gzip-XML as well as ASN.1 PER against EXI is provided in [26], where it is shown that EXI greatly outperforms both other schemes for the used test data set. In [27] Peintner et al. highlight the advantages of schema-enabled EXI in the domain of multimedia applications for embedded systems over the use of plain XML in respect of encoding and decoding performance as well as compression. They especially focus on the encoding of SVG vector graphics and introduce an approach for a more efficient data type representation in combination with EXI in this domain.

To the understanding of the authors at the time of writing this work, there are no previous studies focusing on a quantitative comparison of performance measurements between ASN.1 UPER, protobuf and EXI, specifically on the field of C2X communications. Although, ETSI ITS standards for the facility layer define the encoding mechanisms as ASN.1, this work should provide some insight for the viability of an alternative based on an open source development (i.e., protobuf) or on EXI.

An alternative suggestion to binary encoding of the security envelope using ASN.1 encoding has been proposed in a recent ETSI ITS draft standard [15]. However, there are currently no performance studies available providing insights on which alternative should be selected. Moreover, EXI encoding has not been considered for ETSI ITS data representation so far. Another contribution of this work is to provide some information on the performance comparison of these encoding schemes on different computing platforms such as embedded systems.

III. Performance Requirements and Measurements

The performance metrics considered in the evaluation of the different encoding schemes are

1) computation time,
2) memory footprint on computation and
3) encoded data length.

Aspects 1 and 2 clearly focus on the required computing power for the encoding and decoding process. As ETSI ITS technology shall be implemented in embedded systems, e.g., in vehicles, these criteria are quite important due to the limited resources typically available in such systems. The used tools and methodology to measure these kind of metrics are described in Section IV-B.

The length of the encoded data is a criteria which mostly influences the required communication bandwidth on the wireless channel. It directly determines how long it takes to communicate a data packet over the air. Given that a communication channel has a limited capacity, the length of the encoded messages directly influences the number of possible transmissions over the air in a specific time span. Additionally, ETSI ITS uses only a single control channel to distribute important CAMs and DENMs. Therefore, an increased size of the encoded data packet directly leads to a decrease in system performance and scalability.

IV. Target Platforms

In order to obtain reliable results for our performance study, different hardware platforms with a common software configuration have been used. Details about the used hardware are given in Section IV-A, while the software framework is discussed in Section IV-B.

A. Hardware

To execute our performance measurements of the data serialization schemes in question, we have used three different platforms. The reason is to show the influence of different used hardware technologies as well as to exclude effects on the overall performance study caused by a single processor technology. Table II summarizes the main characteristics of the three platforms used during our experiments.

TABLE II. Used CPU hardware and achievable measurement accuracy via Linux clock counters.

type	AMD Geode LX	Intel Atom Z520PT	Intel Core i7-2640M
clock speed	500 MHz	1.33 GHz	2.8 GHz
clock res.	2 ns	1 ns	1 ns

More details about the used processor technologies can be found in references [28]–[30].

The clock resolution given in Table II was obtained by using the clock_getres() [31] function on the individual platforms in the software environment described in the next section.

B. Software

On all platforms, a standard Debian Linux [32] system with kernel version 3.16.0 was used as the underlying operating system during the performance study. Furthermore, ASN.1 related functionality was provided by the FFASN1 Compiler [33]. Additionally, correctness of encoding as well as the encoded data length was double checked with the ASN.1 library from OSS Nokalva [9]. Protobuf was used in version 2.5.0 as provided by the Debian distribution.

EXI related functionality was provided by the Embeddable EXI Processor in C (exip) library [34]. A double check of the correctness of the encoding and the preparation of the schemes for exip have been done with the Java-based OpenEXI software [35]. For binary encoding of the security envelope the implementation from the ezCar2X framework [36] was used. All used software was compiled on the target with the GCC compiler version 4.8.2 [37]. Thereby, strong optimization was enabled with the *-O3* compiler flag.

In contrast to our prior work in [1], we focus the performance study in this work on the time optimized (TOED) implementation variants of the regarded data (de-)serialization schemes. This is done, as the results in [1] clearly show that computational processing on all used target platforms is bound by pure CPU speed. Thus, the TOED implementations clearly outperform their space optimized (SOED) counterparts in regard to all considered performance criteria. For more details about this aspect the reader is referred to [1].

For timing measurements, the Linux kernel high performance counters have been used, which can be accessed from user space by calling the clock_gettime() function [31]. Thereby, CLOCK_PROCESS_CPUTIME_ID was used as the clock ID in order to determine only the time spent in the process, which contains the algorithm to be measured. An accuracy of up to 1 ns can be achieved, if the underlying hardware permits such accurate measurements [38]. In order to make the measurements more accurate, the suggestions from [39] for avoiding effects of out-of-order execution have been applied. Therefore, the CPUID instruction was executed before and after calling the clock_gettime() function.

The described methodology for time measurements is preferred over directly reading the CPU's time stamp counter (TSC), which is for example used in [39]. The reason is that while [39] uses operations only available inside the Linux kernel, the measurements in our performance study are done in the user space. Therefore, certain prerequisites of the approach from [39] like disabling of interrupts or scheduling cannot be fulfilled. Hence, we rely on the implementation of the clock counter in the Linux kernel.

An algorithm's main memory footprint (heap as well as stack usage) was measured by the help of the so called *malloc_count* framework [40]. This framework allows arbitrary parts of a program to be traced by inserting dedicated function calls into it. These calls where only used during memory measurements and were removed during timing measurements as they would introduce overhead. Other memory tracing tools like *massiv* from the valgrind framework [41] do not allow adjustment of the measurement procedure with such fine granularity. Therefore, malloc_count was used to obtain the results presented in Section V-C.

V. Performance Study

The conducted performance study is discussed in detail in this section. Firstly, the data content to be used for (de-)serialization is described in Section V-A. Secondly, the procedure for generating encoding rules for serialization schemes not present in current standards (i.e., protobuf and EXI) is introduced in Section V-B. Finally, Section V-C provides the results obtained in the extensive performance study, using the methodology from Section IV on content described in Sections V-A and V-B.

A. Content for Encoding and Decoding

For this performance study, we have selected the subset of fields that are defined as mandatory in the standards of CAM and DENM. This implies that the results showed here provide a lower bound of performance of all the considered encoding schemes. For these messages, we have used real data within the message content as far as possible, e.g., we included real time stamps and coordinates.

The studied security envelopes consist of the message fields as specified in [14] and [15], where all defined security profiles (three in total) are taken into consideration. Additionally, for the CAM security profile (security profile number 1) two cases have to be distinguished. The corresponding envelope can hold a signed certificate or just an eight byte hash value of the certificate. Both cases have been included in the performance study.

It should be noted, that the depth of nested data structures significantly affects the performance of encoding and decoding mechanisms. Thus, an overview of the hierarchy of data elements (often called containers in the ETSI ITS context) is given in Table III. Four different cases are considered for the security envelope, which relate to the three different security profiles from [14].

TABLE III. Overview of nesting of data fields for CAM, DENM and security envelope.

nesting level	1	2	3	4	5
CAM	2	4	4	20	10
DENM	4	8	4	0	0
sec. profile 1 (CAM) w/o cert.	5	16	4	0	0
sec. profile 1 (CAM) w. cert.	5	22	21	11	0
sec. profile 2 (DENM)	5	22	23	11	0
sec. profile 3 (Generic)	5	22	22	11	0

The numbers in Table III give the amount of data sets (mandatory and optional) found at the different nesting levels. To obtain the figures in Table III, the full data sets were represented in a tree structure. As we use only mandatory fields in our performance study, the elements of sub-trees following an optional element are not counted. Nesting level one means the top level of the data packet, whereas nesting level five relates to the data elements at the most deeply nested position inside the data packet.

In order to separate the security component tests from others, no real payload was used on these tests. For the case of binary encoding, the envelope only includes the mandatory one byte dummy payload as specified in the standard [14].

Listing 1. VerificationKey element from the security envelope as implemented according to the standard.

```
<s0:VerificationKey>
  <s0:Key>
    <s0:EcdsaNistp256WithSha256>
      <s0:publicKey>
        <s0:CompressedLsbY0>
          <s0:x>FFFFFFFF</s0:x>
        </s0:CompressedLsbY0>
      </s0:publicKey>
    </s0:EcdsaNistp256WithSha256>
  </s0:Key>
</s0:VerificationKey>
```

Listing 2. Optimized VerificationKey element from the security envelope.

```
<s0:SubjectAttributeVerificationKeyEcdsa
NistP256WithSha256CompressedLsbY0>FFFFFFFF
</s0:SubjectAttributeVerificationKeyEcdsa
NistP256WithSha256CompressedLsbY0>
```

B. Encoding Rules for Google Protocol Buffers and Efficient XML Interchange Format

The definition files for protobuf and the XML scheme files for EXI were derived from the ASN.1 definitions given in standards [7], [8], [15]. Transformation from ASN.1 definitions to protobuf and EXI is straightforward due to the low number of available data types in both. During the transformation process always the smallest protobuf (or EXI) data type, which is able to hold the corresponding ASN.1 data type, was selected to avoid introducing unnecessary overhead.

Protobuf does not provide a data element for choices, thus all possible subjects of a choice where chosen to be optional elements. This also means that, the protobuf library does not provide any possibility to check whether exactly one of the to be chosen elements was actually chosen. Thus, this check is left to the user of the auto-generated code.

In the performance study case for EXI, two approaches were followed. At first, a full mapping of the standard to an EXI schema has been developed. These schema files contain a lot of nesting levels, leading often to (informationally) unnecessary content. On the other hand, this makes the existing schemes easy to expand and very structured. However, in all cases the full schema description has to be updated on both sender and receiver when introducing fundamental changes in the message structure. Since one of the key parameters in this study is the size of the encoded messages, we opted for introducing data optimized schemes. In other words, in the schema files the unnecessary nesting levels are merged, thus decreasing the number of options in the EXI grammars. The difference in respect to XML structure of the elements can be seen exemplarily in Listings 1 and 2.

As one can see from comparing Listings 1 and 2, the number of tags required for storing the same amount of payload is reduced from six to only one. Thus, this clearly reduces the amount of metadata in the serialized data, which leads to reduced encoding length.

C. Results of Performance Study

The results of the conducted performance study regarding memory consumption and encoded output length are summarized below. At first, in Tables IV (CAM), V (DENM), and VI (security envelope) results for encoding (i.e., data serialization) are given. Second, results for decoding (i.e., data deserialization) are provided in Tables VIII and IX. In the following, individual results for these message contents are analyzed in detail.

Memory requirements, as well as encoding length are independent of the used CPU architecture. Therefore, just a single result is given for these criteria in the following analysis. Runtime performance, which is clearly processor specific, is presented later on.

In the following analysis, all encoding lengths and memory consumption measurement results are given in bytes. To provide a fair comparison, all memory consumption and runtime measurements for the EXI data serialization scheme give the results for normal, i.e., non optimized, data representation. At first, the results for data encoding, i.e., serialization into the data format transmitted over the network, are given in Section V-C1. A discussion about the results for data decoding, i.e., deserialization of data received from the network follows in Section V-C2.

1) Data Encoding: Data encoding (i.e., serialization) is studied in the following, as it happens at the sender side of a communication connection.

At first, encoding performance for CAMs is studied in detail. The achieved results are summarized in Table IV. Thereby, the value in brackets in the EXI column gives the achieved message size for the case of using the optimized message definition set as introduced in Section V-B.

TABLE IV. ENCODING PERFORMANCE RESULTS FOR CAMS.

	protobuf	ASN.1	EXI
heap / stack	242 / 1864	66 / 3112	62656 / 210
encoded length	165	41	64 (opt: 61)

From Table IV, one can see clearly that protobuf generates almost four times more output bytes than ASN.1 for an encoded CAM. Both the standard as well as the optimization variants of EXI encoding are clearly outperformed by ASN.1, but achieve smaller encoding size than protobuf.

The generated protobuf code uses less memory (cumulative heap and stack) than the ASN.1 implementation. From the measurement results, it is clear that both protobuf and ASN.1, outperform the EXI implementation in regard to memory usage. This is because, the chosen EXI implementation does not use any static a-priori knowledge like the auto-generated code used for ASN.1 and protobuf. Instead, the used library builds up all required trees for encoding on demand in memory. This clearly leads to increased memory consumption and runtime, as one can see from runtime measurement results given in the following. Thus, for a production system one would choose to use a less flexible, but more memory and runtime efficient implementation.

We now study the encoding performance of DENMs. The corresponding results are given in Table V.

TABLE V. ENCODING PERFORMANCE RESULTS FOR DENMS.

enc. type	protobuf	ASN.1	EXI
heap / stack	126 / 1752	75 / 2792	61608 / 175
encoded length	114	43	52 (opt: 51)

As one can clearly see, the memory consumption is similar to the encoding of CAMs but somewhat lower. This is in line with the smaller size of encoded data. As less data has to be encoded, a lower memory consumption can be expected. Additionally, the protobuf encoding shows again the smallest memory footprint of all of the shown four encoding schemes. Furthermore, protobuf performs worst in encoded length, however it only needs roughly three times as much space as ASN.1 compared to almost four times for CAMs.

Moreover, the difference between encoding lengths for ASN.1 and EXI in Table V is significantly smaller than for the CAM case. From studying the different encoding rules, one can see that protobuf introduces more overhead for deep nesting structures than ASN.1 does. From the data analysis provided in Table III, one can see that CAM uses much more and much deeper nested data structures compared to their counterparts in DENM and security envelope. Thus, the difference between the overhead caused by protobuf for the CAM and DENM data sets is as can be expected.

Table VI gives the performance results for main memory consumption as well as encoding length for the ETSI ITS security envelope.

TABLE VI. ENCODING PERFORMANCE RESULTS FOR THE SECURITY ENVELOPE.

enc. type	profile	heap/stack	enc. length
binary	1 no cert.	240 / 12168	96
	1 cert.	798 / 15800	222
	2	798 / 15800	233
	3	798 / 15800	230
protobuf	1 no cert.	1784 / 13528	133
	1 cert.	3819 / 15016	306
	2	4023 / 15016	318
	3	3865 / 15016	312
ASN.1	1 no cert.	1463 / 19784	88
	1 cert.	2186 / 20528	240
	2	2186 / 20528	249
	3	2186 / 20528	247
EXI	1 no cert.	61760 / 680	90 (opt: 87)
	1 cert.	63313 / 680	210 (opt: 201)
	2	63553 / 680	215 (opt: 206)
	3	63457 / 680	213 (opt: 204)

In Table VI the profile column gives the number of the applied security profile as defined in [14]. As described above in Section V-A, the two cases of an envelope with and without certificate have to be distinguished for security profile number one (used for CAMs).

Comparing the overhead introduced by protobuf encoding, its size is between the overheads for CAM (being larger) and the one for DENM (being smaller). Such behavior can be expected, as the nesting of the security envelope is deeper than the one for DENM, but no such deep as for CAM (see also Table III).

The encoding lengths for security profiles two and three are only different for the case of binary encoding and not for ASN.1 encoding, as the data field called *message type* is optional according to [14] but required according to the ASN.1

definition given in [15]. As the only difference between these two security profiles is the presence of the message type data field, this difference vanishes in the case of ASN.1 encoding. Therefore, memory consumption is identical for these two cases. In order for a difference between the two security profiles to exist, our protobuf and EXI definitions declare the message type field as being optional. From a semantical perspective, it makes no sense to give a message type in case of profile number three, as this is the default profile for messages of type *Generic* [14].

One can see from the most right column of Table VI that, in all cases binary encoding clearly outperforms protobuf in respect to achieved encoding length. Additionally, it outperforms ASN.1 encoding in three out of four cases, the only exception being the case of security profile number 1 without certificate. In this case, ASN.1 encoding uses only 9 bytes less than binary encoding. However, for the case with certificate and security profile one, ASN.1 requires 19 more bytes than binary encoding. Furthermore, binary encoding requires 18 bytes less for security profile number two against ASN.1 and 21 bytes less for security profile number 3, respectively.

The results from Table VI clearly show that the normal EXI encoding scheme achieves the smallest packet size for security profile one with certificate as well as profiles two and three. Additionally, it outperforms binary and protobuf encoding for the case of security profile one without certificate and is only slightly outperformed by the ASN.1 encoding scheme. However, the optimized variant of EXI encoding significantly outperforms all other schemes in regard to message size.

In order to decide which encoding scheme performs best for security profile one, the average size of the security envelope should be considered. Due to varying CAM emission frequency (from 1 to 10 Hz) and the different certificate inclusion rules (see [14]), only a lower limit for the average size of the security envelope for profile one can be given. The average size of the security envelope $\overline{s}_{sec}$ is to be calculated by

$$\overline{s}_{sec} = \frac{(f_{CAM} - f_{cert}) \cdot s_{w/o} + f_{cert} \cdot s_w}{f_{CAM}}; f_{cert} \leq f_{CAM}$$

Thereby, the size of the security envelope without certificate is denoted by $s_{w/o}$ and the one with included certificate by s_w. To calculate the metric of the lower limit of $\overline{s}_{sec}$, the maximum CAM emission frequency f_{CAM}, of 10 Hz, together with the minimum certificate inclusion frequency f_{cert}, of 1 Hz, is used. The different values of this metric for the regarded encoding schemes are given in Table VII.

TABLE VII. MINIMUM AVERAGE SIZE OF THE SECURITY ENVELOPE FOR CAMS (SECURITY PROFILE ONE).

encoding scheme	binary	protobuf	ASN.1	EXI
$\min(\overline{s}_{sec})$	108.6	150.3	103.2	102 (opt: 98.4)

One can see from the results provided in Table VII that EXI encoding achieves the best minimum average encoding length. This means, that with EXI encoding the average message size will always be smaller than the one for other encoding schemes, whatever CAM generation rate and certificate inclusion rates are applied. The saved message size for the optimized variant of EXI in comparison to the normal EXI encoding is an additional 3.53%. In comparison to the standardized binary encoding scheme, it even saves the significant amount of 9.39% in message size.

To obtain results for the computation time we ran the measurement procedure described in Section IV-B 10,000 times and computed the average of the measured outcome. Corresponding results for all processor types from Table II are shown in Figures 1, 2, and 3. Please note that the vertical axis of the graphs uses a logarithmic scale. Additionally, for binary encoding only four runtime measurement results are provided per processor as this scheme is not defined for encoding of CAMs and DENMs. Therefore, only the four different kinds of security envelope encoding have been measured.

An overview about the achieved runtime performance measurements on an Intel Core i7 processor is provided in Figure 1 (see also Table II).

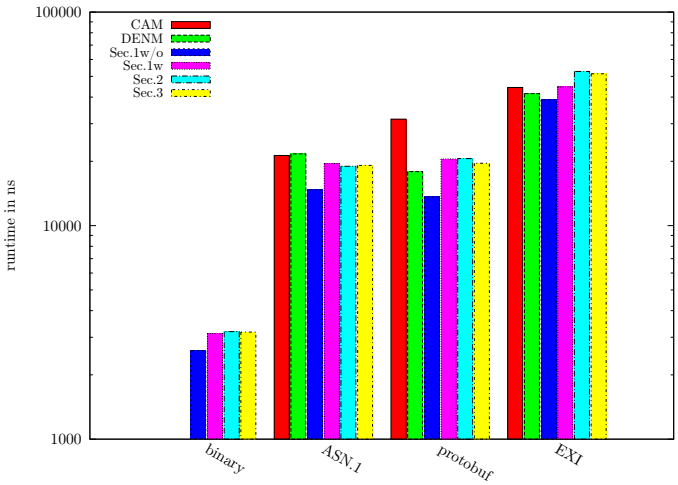


Figure 1. Encoding runtime performance of ETSI ITS CAM, DENM and security envelope encoding on an Intel Core i7 processor.

The obtained results clearly show that, for the security envelope, binary encoding is significantly faster than the two other encoding schemes. Additionally, ASN.1 encoding outperforms its protobuf and EXI counterparts.

A significant source of influence on runtime performance for encoding the security envelope is the high number of small and deeply nested data fields used for defining the security envelope (see also Table III). The achieved results depicted in Figure 1 indicate that binary encoding can handle this kind of structure better than the other encoding schemes can do. Moreover, ASN.1 and protobuf are almost on par and both clearly outperform the EXI mechanism.

To avoid overloading the figures, the computed standard deviation of the measured runtimes are not shown. In general the standard deviation was quite low, e.g., a value of 152 ns was found for binary encoding of the security envelope with security profile one without included certificate. The differences between the obtained results of different encoding schemes for same encoded data content are always bigger than three times the standard deviation of the corresponding runtimes. Therefore, the achieved measurement results can be regarded as reliable.

The results obtained from the runtime measurements on an Intel Atom processor are depicted in Figure 2 (see also Table II).

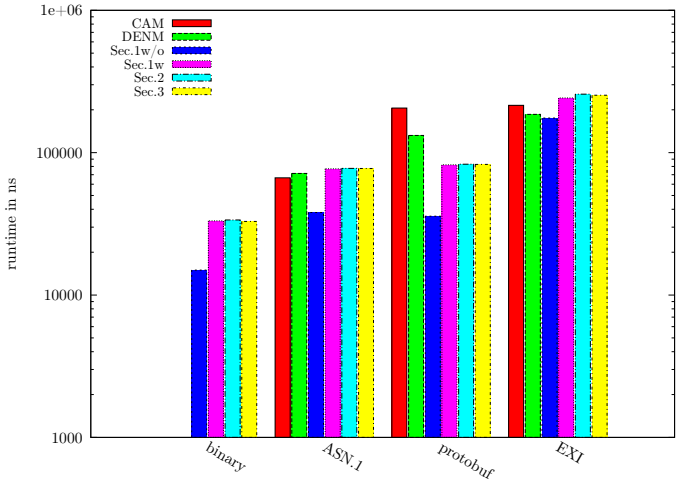


Figure 2. Encoding runtime performance of ETSI ITS CAM, DENM and security envelope encoding on an Intel Atom processor.

Comparing Figure 2 to Figure 1, one can see that except of a general increase in runtime (please note the different scaling of the vertical axis of both figures), the overall results are the same for the Atom and the Core i7 processor technology. Due to the lower processor speed (see also Table II) such an increase in runtime can be expected. However, the increase is somewhat bigger than what can be calculated by just determining the factor one obtains from dividing the respective processor clock speeds. It is reasonable to observe an advantage in the runtime performance of the Core i7, which is due to the improved processor technology such as precaching algorithms, as it was introduced to the market significantly later than the used Atom processor.

Finally, Figure 3 provides the results of runtime measurements conducted using an AMD Geode processor (see also Table II).

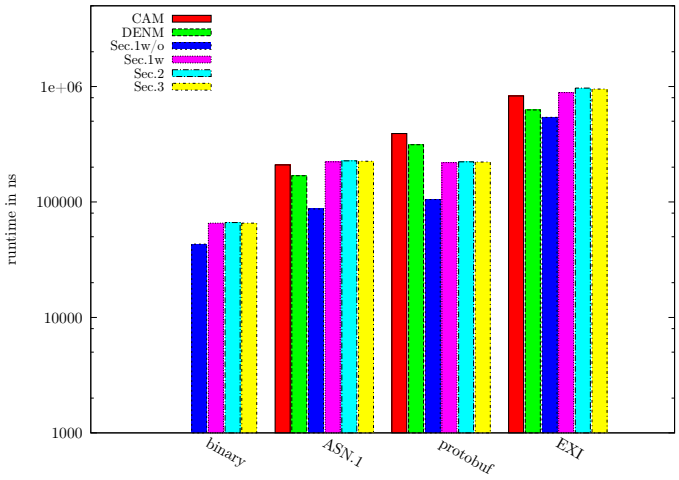


Figure 3. Encoding runtime performance of ETSI ITS CAM, DENM and security envelope encoding on an AMD Geode processor.

From the comparison of results shown in Figure 3 to the results given in Figures 1 and 2, one can see that the overall outcome of the performance study does not change by switching from a modern high speed processor (like the Core i7) to a quite old and low speed processor, like the AMD Geode.

Given the latter statement, we conclude that the achieved results can also be used to interpret the behavior of the studied encoding algorithms within embedded systems using medium speed processors, nevertheless low end, low power processors may possibly behave differently.

In summary, it has been shown that regarding runtime and memory consumption, binary encoding outperforms all other studied encoding schemes running on all platforms. Regarding ASN.1, it can only achieve a shorter encoding length than binary encoding in the case of security profile number 1 without certificate. However, for the security envelope case EXI encoding, especially the optimized variant, outperforms all other serialization schemes in regard to message length.

It is worth to note that, the default timing interval for including a certificate in the security envelope of a CAM is equal to the default sending interval of CAMs (see [14] [7]). The latter means that normally CAMs are sent with a certificate included in the envelope. Therefore, the results show that the newer standard [15] defining the security envelope using ASN.1 significantly deteriorates the performance of its encoding compared to the preceding standard [14], which uses a binary encoding scheme. Furthermore, as ASN.1 does not provide a forward compatibility functionality, like e.g., protobuf would do, there is almost no reason why one should prefer ASN.1 over binary encoding. Instead, one should consider EXI encoding to benefit of a shortening in encoded size of the security envelope of about 9.39%.

The conducted performance study also shows that protobuf cannot be seen as a real alternative to ASN.1 for ETSI ITS data encoding. Protobuf is outperformed by ASN.1 on almost all of the selected important performance criteria on any of the platforms used and for all kinds of data types considered. Protobuf was found to be somewhat smaller compared to the respective ASN.1 counterpart only on the memory footprint parameter for some kinds of data types. Nevertheless, also in those particular cases, protobuf is not able to outperform the binary encoding scheme of the security envelope.

2) Data Decoding: In the following, the results for data decoding (i.e., deserialization) are provided. These mechanisms are required at the receiver side of a message exchange. As the number of other ITS-Ss is typically quite high in a VANET and the majority of messages is broadcast traffic (e.g., CAMs and DENMs), message decoding happens much more often than message encoding. Thus, poor computational performance of a data representation scheme leads to significantly higher penalty at the receiver's side compared to the sender's side.

Table VIII provides an overview of the memory consumption regarding stack and heap usage of the different deserialization mechanisms for CAM and DENM data types.

The obtained results clearly show, that protobuf uses the least amount of memory (cumulative stack and heap) and EXI performs worst in regard to this criteria. Like for encoding

TABLE VIII. MEMORY RELATED DECODING PERFORMANCE RESULTS FOR CAMS AND DENMS.

	protobuf	ASN.1	EXI
CAM: heap / stack	242 / 1800	370 / 2968	3850 / 210
DENM: heap / stack	126 / 1624	816 / 2872	3630 / 135

results (see Tables IV and V), the EXI decoder uses the majority of its memory on the heap in contrast to both protobuf and ASN.1, which take the majority of their memory from the stack.

Table IX summarizes the results for memory usage of the different deserialization schemes for the security envelope.

TABLE IX. DECODING PERFORMANCE RESULTS FOR THE SECURITY ENVELOPE.

enc. type	profile	heap/stack
binary	1 no cert.	872 / 15480
	1 cert.	1709 / 19208
	2	1773 / 19208
	3	1717 / 19208
protobuf	1 no cert.	1916 / 19992
	1 cert.	3665 / 20632
	2	3869 / 20632
	3	3711 / 20632
ASN.1	1 no cert.	1296 / 13016
	1 cert.	4255 / 14040
	2	4327 / 14040
	3	4311 / 14040
EXI	1 no cert.	13375 / 1080
	1 cert.	14131 / 1100
	2	14195 / 1140
	3	14198 / 1136

In contrast to results for CAM and DEMN deserialization, for decoding of the security envelope ASN.1 uses less memory than protobuf, which showed the best performance for CAM as well as DENM. Furthermore, EXI exhibits the smallest memory footprint for all security profiles, significantly outperforming binary and ASN.1 decoding schemes. This clearly shows the dependence of a data (de-)serialization scheme on the used structure of the message.

Memory usage of EXI decoding is much smaller compared to encoding (see also Table VI). This is because the chosen decoder design does not try to build a full message tree in memory before returning the decoded message to the user. Instead, the approach is more like the one for simple binary decoding. The data packet is parsed element by element and for each primitive data type found (e.g., an integer) an a-priori registered callback function (provided by the user) is called. This usage of a-priori information clearly reduces memory consumption inside the decoding method significantly.

Runtime measurements of the different message decoding mechanisms on the three regarded hardware platforms are discussed in the following.

Figure 4 provides the results of runtime measurements on the Core i7 platform.

As one can see from Figure 4, binary encoding significantly outperforms its counterparts for all variants of the security envelope. However, for decoding the gap between binary and ASN.1 is smaller than for encoding (see also Figure 1).

Additionally, ASN.1 performs best for both CAM and DENM decoding. An interesting finding is that the performance gap between ASN.1 and protobuf as well as EXI

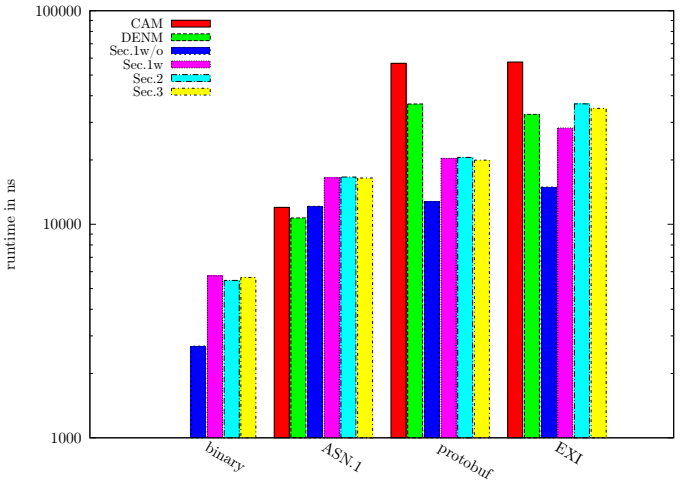


Figure 4. Decoding runtime performance of ETSI ITS CAM, DENM and security envelope encoding on an Intel Core i7 processor.

for CAM and DENM is much bigger than for the security envelope. What is more, protobuf performs poorly for the case of CAM and DENM. EXI shows the worst decoding runtime performance for the security envelope. Furthermore, for protobuf and EXI, decoding of CAM and DENM takes longer than decoding of a security envelope with profile one with certificate. This is a quite unexpected result, as the envelope is significantly longer than a CAM or DENM (see also Tables IV, V, and VI).

Comparison of the decoding results from Figure 4 with encoding results from Figure 1 shows that for all data representation mechanisms decoding is faster than encoding. This is clearly a beneficial property for VANETs, as the number of received (i.e., decoded) messages will typically greatly outnumber the number of sent (i.e., encoded) messages.

Figure 5 gives the runtime measurement results for the Intel Atom platform.

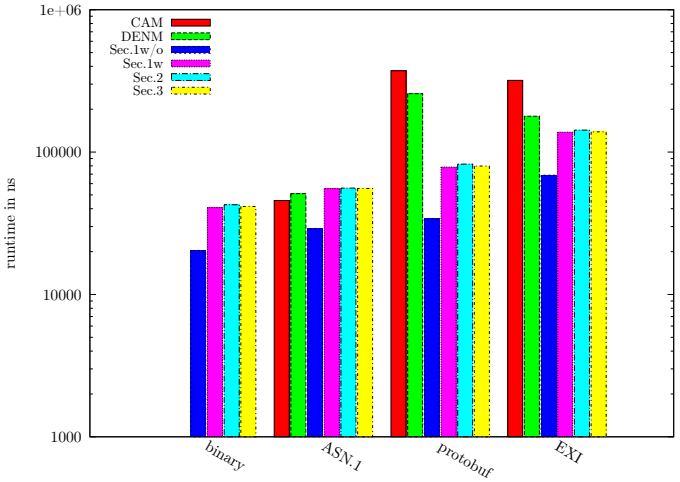


Figure 5. Decoding runtime performance of ETSI ITS CAM, DENM and security envelope encoding on an Intel Atom processor.

While the overall amount of required runtime increases in comparison to the results for the Core i7 platform, no

significant change in the relationship between the different deserialization methods can be obtained.

Finally, Figure 6 provides runtime measurement results for the AMD Geode platform.

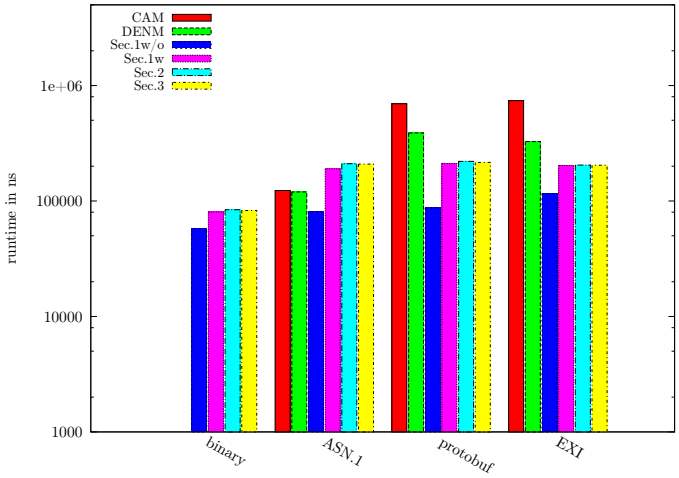


Figure 6. Decoding runtime performance of ETSI ITS CAM, DENM and security envelope encoding on an AMD Geode processor.

As before, it can be seen that the overall processing time scales again when compared to the Atom and Core i7 platforms, but in general the relation between deserialization mechanisms is the same as with the other platforms. Thus, it can be concluded that the obtained results for the relation between processing speeds of the different data representation schemes is the same on all regarded platforms.

In summary, the obtained performance measurement results for data deserialization are in line with the results of data serialization provided in Section V-C1.

An overall conclusion about the achieved results in this work is to be found in the following Section VI.

VI. Conclusion and Future Work

Efficient data encoding schemes are required for future bandwidth-limited C2X communication. In this work, we have addressed three main performance metrics in C2X communications, which are encoded data length, runtime and memory footprint. A study on these metrics for the ASN.1, Google Protocol Buffers (protobuf) and the Efficient XML Interchange format (EXI) encoding schemes has been performed using ETSI ITS CAM and DENM messages as well as for their security envelopes. On the latter, we have further evaluated these metrics for the case of binary encoding, too. To make the study as independent on the hardware as possible, the evaluation was done using three different processor technologies. Our work also presents the used methodology for obtaining the mentioned performance metrics.

The results presented show that the outlined measurement methodology is able to provide the required performance characteristics in a reliable way. Additionally, it was found that the performance of the different encoding technologies is independent of the used processor technology.

From the presented results, it is clear that the performance of protobuf and EXI schemes are almost always outperformed by ASN.1 encoding w.r.t. the required encoding delay or runtime. Only in a minor amount of the studied cases, protobuf outperformed ASN.1 encoding with regard to its memory footprint. EXI showed to be the most expensive scheme in terms of memory footprint. In terms of encoding length for the cases of CAM and DENM ASN.1 UPER encoding performs better compared to EXI and protobuf. For these cases, the differences in length between the serialized information generated by EXI schemes and ASN.1 were considerable but small. In contrast, protobuf serialized message lengths are so large that this scheme cannot be used in C2X communication.

An important result of the conducted performance study is that binary encoding greatly outperforms ASN.1 encoding in the clear majority of cases for the security envelope. ASN.1 actually outperformed its binary counterpart with respect to encoded data length only in one of the studied cases (security envelope for CAMs without certificate). Regarding runtime, binary encoding performs significantly better in all studied cases. The latter implies that the recent shift from binary towards ASN.1 encoding (from [14] to [15]) is not justified at least by the mentioned performance metrics.

In case a more compact representation of the security envelope is required than binary encoding can provide, one should consider to move to EXI data representation instead of its ASN.1 counterpart. The EXI variant always provides a smaller serialization size in average and current performance burdens are likely to be overcome with an implementation being more targeted to the specific ETSI ITS use case. Therefore, the authors propose to conduct additional studies involving either extensive simulations or field tests using both technologies before finalizing the corresponding standard in order to determine which encoding scheme should be used for mass rollout of the future ETSI ITS system.

Directions on future work may include an extension of the provided performance study regarding new upcoming platform independent encoding schemes like Apache Avro [17]. Such systems may provide more flexibility regarding how to organize the encoded data. However, future research has to show whether these improvements have to be paid for by a performance degradation limiting practical usability. Additionally, more runtime efficient implementations of the used Efficient XML Interchange format can be studied to enhance practical usability of this data representation scheme.

References

[1] S. Bittl, A. A. Gonzalez, and W. Heidrich, “Performance Comparision of Encoding Schemes for ETSI ITS C2X Communication Systems,” in Third International Conference on Advances in Vehicular Systems, Technologies and Applications, June 2014, pp. 58–63.

[2] “Memorandum of Understanding for OEMs within the CAR 2 CAR Communication Consortium on Deployment Strategy for cooperative ITS in Europe,” June 2011, v 4.0102.

[3] J. Harding, G. R. Powell, R. F. Yoon, J., C. Doyle, D. Sade, M. Lukuc, J. Simons, and J. Wang, “Vehicle-to-Vehicle Communications: Readiness of V2V Technology for Application,” Washington, DC: National Highway Traffic Safety Administration, Tech. Rep. DOT HS 812 014, Aug. 2014.

[4] Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Std., Rev. 5th, Nov. 2008.

[5] D. Crockford, "The application/json Media Type for JavaScript Object Notation (JSON)," Network Working Group, IETF, RFC 4627, July 2006.

[6] Intelligent Transport Systems (ITS); European profile standard for the physical and medium access control layer of Intelligent Transport Systems operating in the 5 GHz frequency band, ETSI European Standard 202 663, Rev. V1.1.0.

[7] Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service, ETSI European Standard 302 637-2, Rev. V1.3.0, Aug. 2013.

[8] Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service, ETSI European Standard 302 637-3, Rev. V1.2.0, Aug. 2013.

[9] OSS Nokalva, Inc, "ASN.1 Tools for C Overview," online: http://www.oss.com/asn1/products/asn1-c/asn1-c.html, Jan. 2014, retrieved: 03.2015.

[10] Google, "Protocol Buffers - Google Developers," online https://developers.google.com/protocol-buffers/, Apr. 2012, retrieved: 03.2015.

[11] ——, "Protocol Buffers. Googles Data Interchange Format." online http://code.google.com/p/protobuf/, Jan. 2014, retrieved: 03.2015.

[12] Efficient XML Interchange (EXI) Format, W3C Std. 1.0, Rev. 2nd, Feb. 2014.

[13] Road vehicles – Vehicle-to-Grid Communication Interface – Part 2: Network and application protocol requirements, ISO Std. ISO 15 118-2:2014, Mar. 2014.

[14] Intelligent Transport Systems (ITS); Security; Security header and certificate formats, ETSI Technical Specification 103 097, Rev. V1.1.1.

[15] Intelligent Transport Systems (ITS); Security; Security header and certificate formats, ETSI Technical Specification 103 097, Rev. V2.1.1.

[16] Google, "Encoding - Protocol Buffers - Google Developers," online https://developers.google.com/protocol-buffers/docs/encoding, Sep. 2014, retrieved: 03.2015.

[17] J. Russell and R. Cohn, Apache Avro. Book on Demand, 2012.

[18] L. M. Surhone, M. T. Tennoe, and S. F. Henssonow, Apache Thrift. VDM Publishing, 2010.

[19] S. Furuhashi, "MessagePack: It's like JSON, but fast and small," online: http://msgpack.org/, Jan. 2014, retrieved: 03.2015.

[20] D. Gupta, "Thrift vs. Protocol Buffers," online: http://old.floatingsun.net/articles/thrift-vs-protocol-buffers/, May 2011, retrieved: 05.2014.

[21] OSS Nokalva, Inc, "Alternative Binary Representations of the XML Information Set based on ASN.1," online: www.w3.org/2003/08/binary-interchange-workshop/32-OSS-Nokalva-Position-Paper-updated.pdf, Aug. 2013, retrieved: 03.2015.

[22] M. C. Smith, "Comparing the Performance of Abstract Syntax Notation One (ASN.1) vs eXtensible Markup Language (XML)," in Proceedings of the Terena Networking Conference, 2003.

[23] "Using Internet data in Android Applications," online: http://www.ibm.com/developerworks/xml/library/x-dataAndroid/x-dataAndroid-pdf.pdf, June 2010, accessed: February 27th, 2014.

[24] "The gzip homepage," online: http://www.gzip.org, July 2003, accessed: February 27th, 2014.

[25] B. Gil and P. Trezentos, "Impacts of data interchange formats on energy consumption and performance in smartphones," in Proceedings of the 2011 Workshop on Open Source and Design of Communication, 2011, pp. 1–6.

[26] C. Bournez, "Efficient XML Interchange Evaluation," W3C, Tech. Rep., Apr. 2009.

[27] D. Peintner, H. Kosch, and J. Heuer, "Efficient XML Interchange for Rich Internet Applications," in IEEE International Conference on Multimedia and Expo, June 2009, pp. 149–152.

[28] 2nd Generation Intel Core Processor Family, Datasheet, Vol.1, 8th ed., Intel, June 2013, doc. No. 324641-008.

[29] Intel Atom Processor Z5XX Series, Datasheet, 3rd ed., Intel, June 2010, doc. No. 319535-003US.

[30] AMD Geode LX Processor Family, AMD, Feb. 2014, doc. No. 33358E.

[31] ISO, "ISO/IEC 9945:2008 Information technology – Portable Operating System Interface (POSIX®)," May 2009, International Organization for Standardization, Geneva, Switzerland.

[32] "Debian – The Universal Operating System," online: http://www.debian.org/, Dez. 2013, retrieved: 05.2014.

[33] F. Bellard, "FFASN1 Compiler," online: http://bellard.org/ffasn1/, Sept. 2012, accessed: 31.01.2015.

[34] R. Kyusakov, "Embeddable EXI Processor in C," http://exip.sourceforge.net/, Nov. 2014, retrieved: 03.2015.

[35] Fujitsu Laboratories of America, "OpenEXI - EXI implementations in Java and C#," https://sourceforge.net/projects/openexi/, Jan. 2015, retrieved: 03.2015.

[36] Fraunhofer ESK, "ezCar2X: Streamlining application development for networked vehicles," online: http://www.esk.fraunhofer.de/en/projects/ezCar2X.html, Feb. 2014, retrieved: 03.2015.

[37] R. M. Stallman and the GCC Developer Community, Using the GNU Compiler Collection, For GCC version 4.8.2, Free Software Foundation, Oct. 2013.

[38] M. T. Jones, "Kernel APIs, Part 3: Timers and lists in the 2.6 kernel," online: http://www.ibm.com/developerworks/library/l-timers-list/, Mar. 2010.

[39] G. Paoloni, "How to Benchmark Code Execution Times on Intel IA-32 and IA-64 Instruction Set Architectures," Intel, White Paper 324264-001, Sept. 2010.

[40] T. Bingmann, "malloc_count - Tools for Runtime Memory Usage Analysis and Profiling," online: http://panthema.net/2013/malloc_count/, Mar. 2013, retrieved: 05.2014.

[41] J. Seward, N. Nethercote, J. Weidendorfer, and V. D. Team, Valgrind 3.3, 1st ed. Network Theory Ltd., May 2008.

The Role of QoS in WebRTC and IMS-based IPTV Services

Michael Maruschke
Hochschule fuer Telekommunikation Leipzig (HfTL)
University of Applied Sciences,
Leipzig, Germany
Email: maruschke@hftl.de

Jens Zimmermann
Fixed Mobile Engineering Deutschland
Deutsche Telekom Technik GmbH,
Darmstadt, Germany
Email: jzimmermann@telekom.de

Kay Haensge
Telekom Innovation Laboratories
Deutsche Telekom AG
Berlin, Germany
Email:kay.haensge@telekom.de

Tilmann Bach
Technische Planung und Rollout
Deutsche Telekom Technik GmbH,
Berlin, Germany
Email: tilmann.bach@telekom.de

***Abstract*—This paper describes the considerable role of Quality of Service (QoS) for Web Real-Time Communication (WebRTC) clients connected with an IP Multimedia Subsystem (IMS)-based IP Television (IPTV) infrastructure. To raise the quality of experience for IPTV customers, the article focuses on the merging of the technical capabilities arising from both the IMS-based telecommunication networks including IPTV specific components and the WebRTC clients. The ongoing WebRTC standardization process as well as the state of the art WebRTC-QoS trends are considered. To enrich typical IPTV services with appropriate network QoS characteristics a scheme has been developed. The author's concept presents a proposal of an architecture featuring an integrated QoS functionality for WebRTC in conjunction with IPTV services. With our new approach, a WebRTC user inside a 4G mobile network can benefit from the integrated end-to-end quality for real-time IPTV services like Live TV. Composed of several open-source-based testbed solutions, a first prototype has been developed illustrating the QoS initiation procedures primarily.**

Keywords-QoS; WebRTC; IMS-based IPTV; EPS;

I. INTRODUCTION

In times of ever-growing bandwidth needs by Internet users, applications and tightened network resources on side of the network infrastructure providers the importance of QoS mechanics rises heavily. Technologies enabling QoS needs to get deployed more and more corresponding to the communication context (e.g., for conversational voice: delay sensitive, for file transfer: packet loss sensitive). This especially embraces those applications that are not affected by QoS reservations a network provider manages thus far, the so called Over The Top (OTT) services. This includes all the currently established WebRTC services. If the WebRTC client requests for QoS ensured network resources while starting a new communication session, the telecommunications network can provide adequate resources. The advantages are obvious: both end-users and network providers can benefit from such an approach. For the end-users it is possible to experience a high quality even in OTT applications like WebRTC services and for the network provider new business cases are revealed when it is possible to sell the QoS features (inherited from Evolved Packet System (EPS) and fixed line Next Generation Network (NGN) networks) to OTT applications and end-users.

The provisioning of bandwidth many times over the needful proportion (the so called "'overprovisioning"') is going to get increasingly unsuitable. This applies to all kinds of IPTV services, especially if they deliver their content with High Definition (HD) or Ultra-High-Definition resolutions. The realisation of live TV with 4K display resolution would increase the end-to-end data rate enormously. Taking those trends into account, it seems reasonable for network infrastructure operators to provide the bandwidth in a more effective manner by using network QoS techniques.

This journal paper discusses the QoS as an important characteristic for real-time-based telecommunication services in general and for the particular field of IMS-based IPTV services. On the basis of the various QoS requirements for different IPTV services (like Live TV, Audio or Video on Demand) a new QoS resource class mapping is developed by the authors. Based on [1], we propose an architectural concept to enrich an established WebRTC session accessing an IMS-based IPTV network infrastructure with QoS features. Particularly, for an WebRTC end-user, which has access to a 4G mobile network, a new end-to-end QoS control mechanism has been developed and verified. Consuming real-time TV services like Live TV using a Web browser on a mobile device, the user benefits from a QoS concept that combines the well established QoS technology from the 4G mobile network with our new WebRTC QoS enrichment. While the authors paper [1] proposed the principle combination of WebRTC with IMS-based IP-TV services, this journal contribution is focussed on the enrichment of QoS for an WebRTC client consuming TV services with real-time characteristic. In contrast to this, our other publication [2] is aimed to establish QoS only for an conversational voice call. The authors' QoS extension principle is new and neither proposed by established international standardization bodies

nor solved in a practical manner up to now. To verify our proposal, a first testbed has been implemented. Particularly, the QoS activation procedures have been tested.

The present journal paper is structured as follows: Specifying the used terminology like RTC and QoS, Section II offers a survey of the current status of QoS in context of WebRTC, of real-time communication in fixed and mobile telecommunication networks and of IPTV services. Based on an IPTV-QoS parameter mapping approach, Section III describes the authors new concept to integrate QoS dynamically in the WebRTC client accessing IMS-based IPTV architecture. (4G mobile network). The new architecture and their specifics are considered and the proof of concept incorporating an 4G mobile network is presented. In Section IV, a conclusion summarises the achieved results of this contribution and gives an perspective.

II. Status Quo

A. Real Time Communication and QoS

Real Time Communication (RTC) is generally characterised by the so called "Real-Time" condition. That implies that the value of the communication depends significantly upon the time at which the data is arriving at the data sink [3]. The throughput time of the data (flow rate) across the network delays the delivery of the data packets to the recipient. The tolerable delay time or latency depends on the type of the desired communication (e.g., conversational audio and video or real-time gaming). Besides the delay time, two other characteristic performance aspects can significantly influence the quality of the real-time communication. First, the circumstance that the transmitted data packets are routed through the network passing an unequal number of network elements, which results in variable arrival times at the recipients side. In packet-based networks, the varying delay of the transferred data is called packet delay variation or simply Delay Variation (DV) [4]. Sometimes, the different packet delivery times are also named as jitter [5]. Furthermore, if data packets reach the incoming data buffer on the receiver side too late, they will be discarded and consequently counted as lost packets.

Basically, QoS encompasses both the service categorization and the overall performance of the network communication for each service category. The International Telecommunication Union - Telecommunication Sector (ITU-T) describes QoS as the unity of all characteristics of a telecommunication service which are necessary to satisfy the service user. This includes the defined and the implied requirements for the complete customer satisfaction. Mean Opinion Score (MOS) methods are often used to indicate the measured or detected service quality. Initially developed for a subjective quality evaluation, the MOS method with it's rating score from 1 (bad) to 5 (excellent) is more and more used for objective QoS measurement methods.

Appropriate QoS characteristics like End-to-End-Delay of the transmitted packets, the Packet Delay Variation and the size of the tolerable Packet Loss are required to satisfy the telecommunication customer's expectations. The end-to-end quality of any telecommunication service depends on the performance of all involved components; the technical end-user system as well as each relevant network entity (see Figure 1). Hence, the QoS one-way-delay parameter is influenced by all components. For speech transmission, an one-way-delay (also known as 'mouth-to-ear-delay') of 150 ms is experienced as

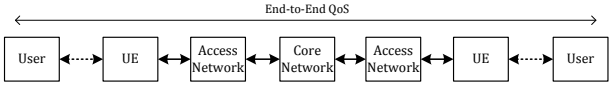


Figure 1: Schematic contributions to end-to-end QoS [7].

very good and all delays above 400 ms are considered to be unacceptable for the consumer [6].

Focusing especially on IP-based networks, several QoS requirements are reasonably for a successful and effective operation of RTC. Therefore, the ITU-T also presents a generic QoS classification for public IP-based networks and the related network performance parameters IP Transfer Delay (IPTD), IP Delay Variation (IPDV), IP Packet Loss Ratio (IPLR) and IP Packet Error Ratio (IPER) (see also table 1 and table 3 of [8]). From a network topology point of view, this recommendation addresses all relevant network parts (Access Network; Aggregation Networks, Core Networks) but not the end-user or home-network side. Table I illustrates the eight defined IP network QoS classes and their corresponding QoS network parameters. Note that the parameter IPER is contributing insignificantly to the overall packet loss and therefore is not shown in the Table I. The value 'U' stands for 'unspecified'.

Furthermore, a guidance on usage of those eight QoS classes is given as follows, outlining corresponding communication examples:

- QoS class 0, for real-time, jitter sensitive and high interactive applications like Voice over IP (VoIP) and Video conferences;
- QoS class 1, for real-time, jitter sensitive and high interactive applications like VoIP and Video conferences, but with less constrained delay requirements;
- QoS class 2, for transaction data, highly interactive (signaling traffic);
- QoS class 3, for transaction data, interactive;
- QoS class 4, for low loss only applications like short transactions, bulk data, non real-time buffered video streaming;
- QoS class 5, for all other traditional applications of default IP network without any QoS demands
- QoS classes 6 and 7 have a provisional character and are designed for applications similar to applications in QoS class 0 or 1, but with more strictly demands for the packet loss rate.

B. QoS in IMS-based Telecommunication Networks

IMS, as an architectural framework for supporting IP multimedia services, was originally designed by the 3rd Generation Partnership Project (3GPP) for mobile core networks and successive expanded for fixed-line-based core networks [9]. It addresses multiple IP multimedia applications like speech and video communication, shared online whiteboards, telepresence conferences and multicast services. To provide these in a flexible and appropriate manner, telecommunication network operators differentiate their services for the customer regarding to the QoS characteristics [9]. Basically, QoS should be negotiable for

TABLE I: IP network QoS class definitions and network performance parameters by ITU-T Rec. Y.1541.

Network Performance Parameter	QoS Classes							
	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
IP Transfer Delay	100 ms	400 ms	100 ms	400 ms	1 s	U	100 ms	400 ms
IP Delay Variation (Jitter)	50 ms	50 ms	U	U	U	U	50 ms	50 ms
IP Packet Loss Rate	0,1 %	0,1%	0,1%	0,1%	0,1%	U	0,001%	0,001%

IP multimedia sessions and their individual media components (like audio or video) both at the time before establishing a connection as well as during the established connection. QoS related concepts (QoS signaling, QoS resource reservation and allocation, etc.) belongs to this framework inherently due to the fact that IMS describes one technological concept to realise a NGN. Next to other capabilities like mobility of packet-based telecommunication networks, the QoS is a key feature in a NGN as defined by the ITU-T [10].

Mobile networks

QoS has always been considered in the standardization process of mobile networks. Therefore, in Universal Mobile Telecommunication System (UMTS) networks (3G) four QoS classes has been introduced. They are named as follows:

- Conversational Class,
- Streaming Class,
- Interactive Class, and
- Background Class [11].

These classes are also entitling the possible use cases for each of the four types. Conversational is used for real-time audio and video communication. Streaming can be used to stream audio or video data towards the User Equipment (UE) by having a small buffer and non-critical real-time constraints. Interactive is used for general user data transfer such as web browsing and application information exchanges. The last and lowest prioritised class is background. It is used for non time-critical applications, like email polling.

With the specification of EPS and fourth generation (4G) mobile networks, the 3GPP has developed a new, only packet-based core network domain (also known as Evolved Packet Core (EPC)) for both conversational voice/video communication as well as other packet or IP-based applications like public Internet communication. For this modern all-IP network, the 3GPP also defines a new QoS concept involving all relevant EPS network components like the UE, the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and the EPC. For enabling QoS inside the EPS a so called EPS Bearer is used to fulfill all requirements of the media delivery. Figure 2, depicts this EPS Bearer, which is correlated with the overall Bearer-based End-to-End (E2E) QoS concept of the 3GPP network infrastructure. The detailed description can be found in [12]. QoS classes are composed of a subset of standardized characteristics, they describe the packet forwarding treatment in 4G networks. For the QoS class determination the so called QoS Class Identifier (QCI) was introduced. Based on typically applications, various QoS groups separated in nine QCI values are defined in [13].

Since the 3GPP standardization institution provides their own QoS treatments, a mapping is always required. Therefore, Table II shows different QoS classes originated in 3GPP Release 99 networks (UMTS-3G) and EPS-based 4G networks, in relation to the main relevant QoS parameters Delay and Packet Loss. It should be clarified that in Table II with ”Prio” the

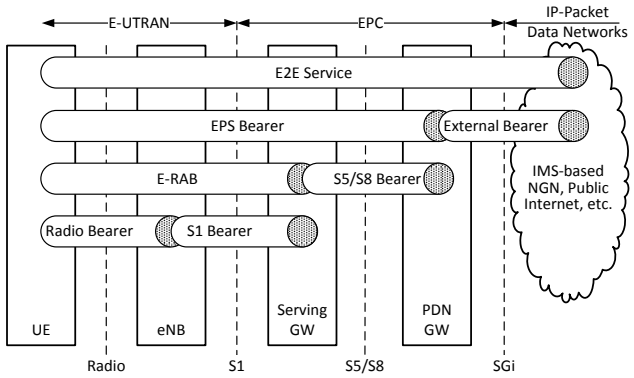


Figure 2: 3GPP overall QoS concept, based on [12].

packet scheduling priority is meant and Packet Error Loss Rate (PELR) stands for an upper bound for the rate of non-congestion-related packet loss. While for the first four QoS classes (QCI 1 to 4) a Guaranteed Bitrate (GBR) is allocated, the following classes (QCI 5 to 9) are supported with Non-GBR characteristics. It is obvious, that real-time applications like audio and video conversation demands more strict packet delay requirements than traditional Internet traffic like file sharing, which is more critical regarding the packet loss characteristic. While for telecommunication carriers the guaranteed delivery of IMS Signaling is relatively important, the customer needs the stricter QoS resources for voice, video and interactive real-time gaming (see also QCI value 3 or 7). It is noticeable that for video streaming services (both live streaming and buffered streaming) exists several QoS relevant QCI values. Thus, live streaming services are typified by QCI value 2 and 7 with similar QoS characteristics (packet delay and packet error loss rate), but with different priority and bitrate guaranties. To handle QoS for buffered streaming services, a couple of QoS classes (QCI values 4, 6, 8 and 9) is specified. The assigned QoS relevant parameters packet delay and packet error loss rate are identical once again meanwhile the priority and the bitrate guarantee differs. Note that QCI value 9 is typically used for the default bearer of a UE for non privileged subscribers.

Fixed line networks

QoS in fixed-line-based IMS networks differs relating to the used IP Connectivity Access Network (IP-CAN). While the 3GGP-based mobile networks use their typical access networks like UMTS Terrestrial Radio Access Network (UTRAN) or E-UTRAN, the fixed-line-based IMS networks provide the telecommunication services to the residential customer via any Digital Subscriber Line (xDSL) technique. Besides the fact that the access transmission technology differs, the

TABLE II: QoS class mapping between 3G and 4G mobile networks, based on [13].

UMTS Traffic Class (3G)	QCI (4G)	Prio	Bitrate Guarantee	Packet Delay Budget	Packet Error Loss Rate	Example Services
Conversational	1	2	GBR	100 ms	1 %	Conversational Voice
Conversational	2	4	GBR	150 ms	0,1 %	Conversational Video (Live Streaming)
Conversational	3	3	GBR	50 ms	0,1 %	Real Time Gaming
Streaming	4	5	GBR	300 ms	0,0001 %	Non-Conversational Video (Buffered Streaming)
Interactive	5	1	Non-GBR	100 ms	0,0001 %	IMS Signaling
Interactive	6	6	Non-GBR	300 ms	0,0001 %	Video (Buffered Streaming), TCP-based (e.g., www, chat, ftp, p2p file sharing, progressive video)
Interactive	7	7	Non-GBR	100 ms	0,1 %	Voice, Video (Live Streaming), Interactive Gaming
Interactive	8	8	Non-GBR	300 ms	0,0001 %	Video (Buffered Streaming), TCP-based (e.g., www, chat, ftp, p2p file sharing, progressive video)
Backround	9	9	Non-GBR	300 ms	0,0001 %	Video (Buffered Streaming), TCP-based (e.g., www, chat, ftp, p2p file sharing, progressive video)

Customer-Premises Equipment (CPE) (any terminal and associated equipment located at a subscriber's premises and connected with a carrier's telecommunication channel) is distinct from the connected network infrastructure. Therefore, mobile network customers are utilizing mobile devices like cellulars, smartphones or tablet PCs. In contrast to this, in fixed line networks the residential customer uses a home-network where various end devices are connected to each other and accessed on the CPE, which can be build by an Integrated Access Device (IAD) for IMS-based NGNs.

In fixed access line NGNs, the QoS management functions are supported by a Resource Admission Control Subsystem (RACS), which is responsible for elements of the policing control including resource reservation and admission control in the access and aggregation networks [14]. Any multimedia services like VoIP or IPTV can request particular QoS parameters, such as data throughput, latency, jitter and packet loss from the transport network side. Then, the RACS is responsible to manage this QoS requests by evaluating this in the context of predefined policy rules, and performing the reservation and allocation of adequate QoS resources through all affected transport network elements.

To ensure QoS aware NGN service delivery in fixed access networks, the RACS specification distinguishes between two abstract QoS architecture principles [14]:

- *guaranteed QoS:* traffic delivery service with absolute demand on some or all of the QoS parameters, such as throughput, latency, jitter and packet loss, and
- *relative QoS:* traffic delivery service without absolute demand on some or all of the QoS parameters.

In contrast to this, the support of QoS unaware ("Best Effort") networks as well as the support of networks that have statically provisioned QoS differentiation does not require any RACS functionality. To determine the various QoS classes the DiffServ classification is applicable [15].

Sometimes, for statically QoS support, the QoS marking inside the IAD-based Home-Network on CPE side is carried out by utilizing the Differentiated Services Code Points (DSCP) classification mechanism for each media flow.

C. QoS in IMS-based IPTV services

In principle, Video Quality is defined as the indicator, which evaluates the quality of the video stream delivered to the user [16]. This indicator describes the perception of the end-user in term of the video quality. Typically, objective perceptual video quality measurements models are utilizing a reference-based approach, like the ITU-T recommendation J.247 [17] describes it. Taking the end-user context by consuming Live TV or Video on Demand (VoD) into account, is it not applicable to operate with a video quality assessment method, which is functional with a full reference. The current issue is that there is no standardized approach for an objective video quality measurement model that does not need any reference signals. Therefore, in context of end-user quality survey of IPTV services the following indicators are proposed to characterize the quality of IPTV services [16]:

- Channel Availability (indicates the availability of Live TV/VoD channels proportional to the attempted channels),
- "Black Screen" Occurrences (e.g., effected by a major loss of video packets during a long period of time),
- Blockiness Occurrences (produced by a low-quality video compression when too few bits are present; it is perceptible by the contrast of color),
- Frozen Picture Occurrences (some picture appearing as stopped/frozen from time to time),
- Lip Desynchronization Occurrences (the synchronization of audio and video stream is not well),
- Zapping Delay (time, which is needed to be switch from one TV/VoD channel to another),
- Transmission Delay (indicates the delay to transmit the audio/video signal from the delivery point to the end-user's TV device; important for some cases like football matches),
- and others.

While reference [16] is focused on the context of end-user quality characterization and their indicators, it is obvious that the network-based QoS parameter Transfer Delay (TD), DV and Packet Loss Rate (PLR) have also high relevance.

As already introduced in [1], IMS-based IPTV services can be differed among each other regarding their kind of service or feature, based on [18]. For instance, linear live Television (TV) or real-time VoD will demand other network performance characteristics then the Electronic Program Guide (EPG) feature. While the linear live TV service demands more stringent QoS performance characteristics like low transfer delay, low packet delay variation and minimal packet loss, the IPTV content control feature EPG accepts less tightened QoS performance. The specification [18] also describes a general approach for dynamic QoS resource modification between Standard Definition (SD)-TV and HD-TV.

To fulfill an adequate QoS support for all typical IPTV services, a high-level guideline for the use of traffic management is given by [19]. Therein, a potential ITU-T Y.1541 performance class mapping for typical IPTV service applications is provided

and visible in Table III. In general, the IPTV service category *Streaming* demands most of the QoS resources, because it is "live"-communication with strict real-time prerequisites. Besides, the rubric *Download* includes video content consuming service like near VoD. It becomes also known as progressive video download principle using by consuming Youtube-Videos.

TABLE III: Potential mapping of IPTV services to ITU-T Y.1541 QoS classes [19].

IPTVservice	Service applications	ITU-T Y.1541 QoS class							
		5	4	3	2	1	0	7	6
Streaming	Live TVcontent						x		x
	Video content					x		x	
	Audio content					x			
	Content control				x				
	Live speech						x		
	Live low- resolution video content						x		
Download	Video content		x						
	Data	x							
Upload	Video content		x						
Message exchange	Interactive			x					
	Non- interactive	x							
Middleware/ application	Portal			x					
	Payment transactions			x					

D. WebRTC and QoS

WebRTC introduction

A main focus of the upcoming WebRTC technology is to integrate real-time communication into standard web browsers without the need of any additional browser plug-in or software. The World Wide Web Consortium (W3C) defines an Application Programming Interface (API) for web developers [20]. It empowers the browser to capture video and audio inputs of the client's device. While the W3C is responsible for the web developer API, the Internet Engineering Task Force (IETF) standardizes all corresponding protocols in an active working group named "'Real-Time Communication in WEB-browsers - RTCweb" [21].

WebRTC has been specified to use secure transport of the Real-Time Transport Protocol (RTP) packets with Secure Real-Time Transport Protocol (SRTP) [22] including the mandatory Datagram Transport Layer Security (DTLS) encryption protocol [23] used for key negotiation [24]. For solving Network Address Translation (NAT) problems, WebRTC also provides Session Traversal Utilities for NAT (STUN) [25], Traversal Using Relays around NAT (TURN) and Interactive Connectivity Establishment (ICE) [26] capabilities. WebRTC requires Session Description Protocol (SDP) for the negotiation of the session properties and uses the whole SDP's Offer/Answer-Model.

The generic architecture of a WebRTC client is described by [27] and illustrated in Figure 3. The components can be described as follows:

- Webserver, which provides the web application to load and includes a server with which the client connects to handle all signaling;
- Browser, a generic web browser;
- Web application, application source code executed by the web browser;
- Browser RTC Function, WebRTC component in the web browser with voice, video and transport engines;
- Signaling Path, which is not specified but is needed to transfer the SDP information between the WebRTC Client and any other signaling endpoint (alternative bypassing the Webserver);
- Media Path, which transports the payload;
- STUN Connection, which is a component to bypass NAT restrictions.

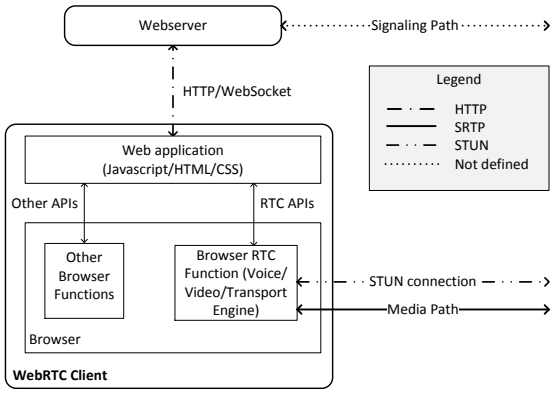


Figure 3: WebRTC client based on [27].

For a WebRTC client to successfully run, it is necessary to use a capable web browser. That means the browser has to implement the Browser RTC Function. Currently, web browsers like Google Chrome, Mozilla Firefox and Opera provide this component by default. Therefore, all devices that are able to run one of these browsers have the ability to use web-based real-time communication. This includes all, desktop and tablet computers, laptops and smartphones. At the moment, there are restrictions in some operation systems like Apple's iOS or Microsoft's Windows Phone.

In this subsection, we intend to find out if QoS is a topic for the developing WebRTC technology and their standardization bodies. Typical use cases for WebRTC and their requirements are described in [28]. Based on a simple video communication service, some use cases are presented, involving voice or video or data communication respectively combinations of those. For instance, to realize a "Multiparty on-line game with voice communication" quick updates of the game state are required, and they have higher priority than the voice [28]. Generally, the browsers should be able to render good audio and video quality for an adequate and acceptable jitter and packet loss values and must support a time synchronized audio and video playback function. If a WebRTC Client is accessed behind a residential router that supports any kind of data traffic prioritization, the user should be able to take advantage of this QoS support, provided by the network side. Summarizing, from an application layer point of view the WebRTC use cases do not define some exact and comparable values for the network related QoS parameter like delay, jitter or packet loss.

DSCP-Marking

An IETF-Draft proposes a QoS mechanism at the WebRTC client side using DSCP [29]. This document provides DSCP values for browsers to use for various classes of traffic. It proposes how WebRTC applications can mark data packets for a packet prioritization. It assumes that residential or wireless networks support traffic preferential treatment, based on DSCP. For all other cases including cellular mobile-based network

access, this suggestion is not appropriate. However, if the real-time packet is transmitted towards a QoS capable core network domain, a QoS class mapping is needed. Besides, client side marking of IP packets is also a topic for the admission control processes. The issue is, if the marking is allowed and the network enforces the requested QoS parameters, other (unwanted) applications may also request prioritized forwarding. This situation may lead to overall high prioritized traffic with no benefit for the actual intended application. Therefore, the authorization of client-side requested QoS needs to be clarified. This draft does not cover any mapping processes for QoS management inside an 3GPP-based mobile network Release 8 or inside a fixed-line-based network, which does not use DSCP marking mechanism.

QoS concepts for WebRTC accessed to a mobile EPS network

Currently, few concepts exist which propose the enrichment of QoS for an WebRTC Client accessing a mobile EPS network. The 3GPP specification (Section Annex U of [9]) proposes a new architecture for including WebRTC Clients into an IMS-based EPC network. Hence, it describes only, that QoS support can be provided. However, any definite method or mechanism for this is missing. To fill this gap, an authors proposal for a QoS support method for WebRTC users, which are connected to an EPS-based IMS network infrastructure is documented in [2]. To fulfill this QoS support, the 3GPP-based architecture (taken from Section Annex U of [9]) became enhanced with the following entities named

- *WebRTC Client with QoS Awareness*,
- *WebRTC QoS Signaling Function (WQSF)*, and
- *eP-CSCF*, which is a modified eP-CSCF*.

This proposed WebRTC QoS Architecture is depicted in Figure 4. The concept and its added entities will be described briefly. Provided in reference [2], the WebRTC communication is enriched with the capabilities to request QoS resources via signaling all used RTC data flows to the WQSF, which adapts the information and starts a signaling session with the Policy and Charging Rules Function (PCRF) of the underlying core access network. This concept reuses the standardized EPS architecture as well as the new proposed integration of WebRTC clients into the IMS network [9]. The WebRTC Client acts as an UE that provides QoS awareness. It allows requesting QoS characteristics during the active conversation phase. With the help of periodical and event-based transmissions of the QoS relevant information (used media type like audio or video, and IP flow information like IP addresses and port numbers) of each established media stream, it is possible to use the EPS related QoS mechanisms dynamically. This means that the web application can start with one media type (e.g., audio) and add another media type (e.g., video) during the conversation phase. All involved components (e.g., end device, network entities) have to support such a dynamic mechanism to enforce the changed QoS resources.

For further understanding of the QoS control mechanisms in the described architecture, Figure 5 depicts the general QoS initiation procedure between the WebRTC Client and the EPS-based network architecture.

The Web Application starts with the QoS control sending relevant application layer QoS requirements, such as flow and media type information, towards the Application Function (AF). After a possible negotiation and session information signaling,

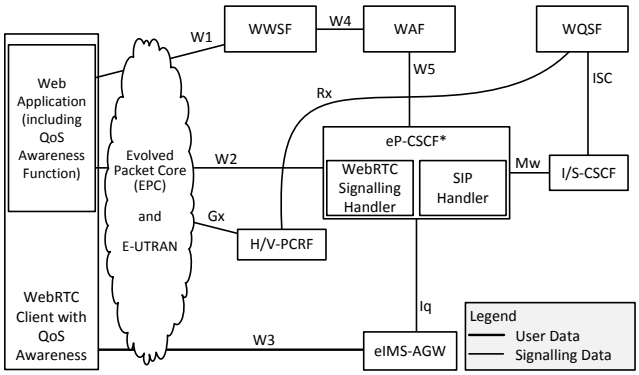


Figure 4: WebRTC QoS architecture for an EPS network [2].

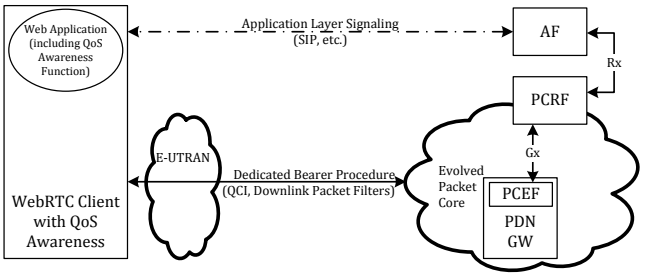


Figure 5: Principle of QoS initation.

the AF is aware of all QoS request parameters and will start the Policy and Charging Control (PCC) procedures by sending the mapped parameters into Diameter Attribute-Value-Pairs (AVP)s via the Rx interface towards the PCRF. The application can rely on the AFs endpoint to request QoS for its application flows. The PCRF will authorize the request and will start the resource allocation via the Gx interface to the Policy Control Enforcement Function (PCEF).

More details including the used QoS signaling and enforcement procedures, the evolved protocols and interfaces respectively message sequence charts are presented in [2].

E. Summary of Status Quo

As we have illustrated in this section, the role of QoS in telecommunication networks is heterogeneous. Depending on the specific scope of the standardization bodies 3GPP and ITU-T, the QoS relevant metrics and parameters differ among themselves. For instance, utilizing the network performance parameters IPTD, IPDV (Jitter) and IPLR the ITU-T recommendation Y.1541 [8] is clearly focused on the IP network layer (Open System Interconnection/International Organization for Standardization (ISO/OSI)-Layer 3 [30]). Regarding the area of applicability of the affected network segments all involved core networks and access networks are intended, but not the UE. Otherwise, the 3GPP defines in their Technical Specification TS 23.203 [13] a QoS parameter called QCI, which is closely associated with packet forwarding treatment characteristics

like guaranty of bitrate, packet scheduling priority, packet delay budget and packet error loss rate. The scope of this standardized QoS parameter QCI is focused on the EPS Bearer lever (ISO/OSI Layer 2) between the mobile UE side, the mobile access network, the mobile core network (e.g., EPC), but not the IP backbone network, which is located behind the mobile core network. Concluding this, it means that a QoS mapping between heterogeneous telecommunication network ecosystems, which are supporting RTC like conversational voice or live TV is not simple. The challenge is to realize end-to-end QoS over such various networks transparently and perceivable for the end-user. From this perspective, QoS should be focused on a *horizontal QoS level* (see also Figure 1). Otherwise, the QoS metrics differ, depending on the functional layer (e.g., the ISO/OSI protocol layers). It is obvious that the overall QoS result depends on close interworking of all involved protocol layers. For instance, in an UE the QoS E2E service can not be realized successful, if all lower QoS layers (EPS Bearer, E-UTRAN Radio Access Bearer (E-RAB) and Radio Bearer) are not assuring their own QoS support (see also Figure 2). This QoS perspective is also called *vertical QoS level*. An overall QoS concept needs to be comply both, the *horizontal* and the *vertical QoS level*.

Independent of the above described QoS taxonomy and metrics, from the end-users perspective the expectations on a communication application are highly relevant. The perceived QoS by the users strongly depends on the performance of the network. However, it is measured by the opinion of the users. A typical subjective metric to measure this QoS performance by the consumers is commonly known as a MOS method. Though various users are consuming the same content (e.g., a Youtube Video) it will result in subjective and different perception and quality ratings. As an example, one user may rate the perceived quality as good, while another user will rate the same communication application as not acceptable.

Regarding IPTV, the standardization bodies define different IPTV Services and for each they assign different objective network QoS classes (see also Table III). For that, in Section III we propose an IPTV Service mapping into the relevant network QoS classification.

III. Concept

A. QoS Enrichment for IMS-based IPTV

The authors concept of the enrichment of QoS for typical IPTV services is based on specifications provided by the standardization bodies ITU-T and European Telecommunications Standards Institute (ETSI) [19] [18]. The most common services and features described are Live/Linear TV or Video, respectively Audio on Demand as streaming application and also Download applications like near video on demand (e.g., Youtube video downloading) or the feature EPG as type of data download.

Various IPTV services with its particular real-time communication characteristics demand different QoS resources. Therein, the IPTV services are related to network QoS classes, defined by 3GPP and ITU-T [13] [8].

We propose an assignment mechanism, which can be used for IPTV services to correlated to their appropriate network QoS classes. This mapping approach is depicted in Table IV.

In this table, a mapping is performed between the IPTV Service Applications, taken from ITU-T Y.1920 with the category '*Example Services*' of EPS QoS classes in 4G networks (see also Table II).

As already described in the summary of Section II, the differences between the two network QoS class approaches (3GPP versus ITU-T) were indicated. Therefore, those consolidation make sense.

As a result on the basis of the correlated IPTV services (see also Table IV), a correlation of concrete services/features to the adequate network QoS classes given by 3GPP and ITU-T is shown. Accordingly, and for instance, the Live/Linear TV service is associated with the EPS QoS class parameter QCI value 2, priority value 3 and GBR, as well as the ITU-T Y.1541 classes 0 or 6. This also encompasses the linked network QoS parameter delay, packet loss and delay variation. Taking into account that the EPS QoS classes and their corresponding parameters Delay Budget and PELR are focused on the network segments UE - Access Network - EPC, on the ISO/OSI protocol layer 2. In contrast to this, the ITU-T QoS classes are aimed at IP core and IP access network parts of the IP layer. Therefore, both the vertical as well as the horizontal QoS levels differ.

As depicted in Table III, the preferred ITU-T QoS classes for video content are class 1 and class 7. Based on this information the mapping to the QCI classes leads to QCI class 4, 6 or 8. Each of these QCI classes share the same values for the technical parameters delay and packet loss. The QCI class 4 is used as favourite, because it has a higher priority as the other two classes and it uses a GBR. For audio content the ITU-T class 1 should be used. The mapping to the QCI approach is similar to video content. As result the QCI class 4 is used for audio content, too.

Another IPTV service, described in [1], is the content control. For this service, the Table III gives information about the use of ITU-T class 2. The mapping of the technical parameters lead to the QCI class 5. Otherwise, compared to real-time audio or video, the content control has not such high demand for the bit rate. But the content control should not be affected by congestion, so a higher priority is more preferable. Therefore, the QCI class 5 is a valid approach.

B. Consolidating EPS and IMS-based IPTV Architectures for QoS support

The proposed architecture is based on the presented IMS-based IPTV architecture from [1], depicted in Figure 6, and the analyzed concept for WebRTC accessing to an mobile EPS network [2]. The principal components are marginally accommodated. The innovation in this approach is the enrichment of QoS for the IMS-based IPTV environment with WebRTC. The consolidated architecture of the proposed concept is depicted in Figure 7. Components and interfaces of the architectures are described in the following section.

Components and interfaces, which result from the approach from [1], are:

- Components:
  - Webserver
  - WebRTC client
  - Signaling GW (SGW)
  - Core IMS
  - SDF
  - SSF
  - SCF
  - MCF
  - MDF (modified for WebRTC)

TABLE IV: Mapping of IPTV services towards 3GPP/ITU-T QoS classes.

IPTV Service Application (ITU-T Y.1920)		IMS-based IPTV Services (ETSI TS 182027)	EPS QoS Classes (3GPP TS.23.203)			QoS Classes (ITU-T Y.1541)
			QCI	Priority	GBR / Non-GBR	
Streaming	Live TV content	Linear / Broadcast TV	2	3	GBR	0; 6
	Video content	VoD, Network PVR, Time-Shift TV	4	5	GBR	1; 7
	Audio content	AoD	4	5	GBR	1
	Content control	Content control	5	1	Non-GBR	2
Download	Video content	Push VoD, Near VoD	6; 8; 9	6; 8; 9	Non-GBR	4
	Data	EPG	9	9	Non-GBR	5
Upload	Video content	Interactive TV	6; 8; 9	6; 8; 9	Non-GBR	4

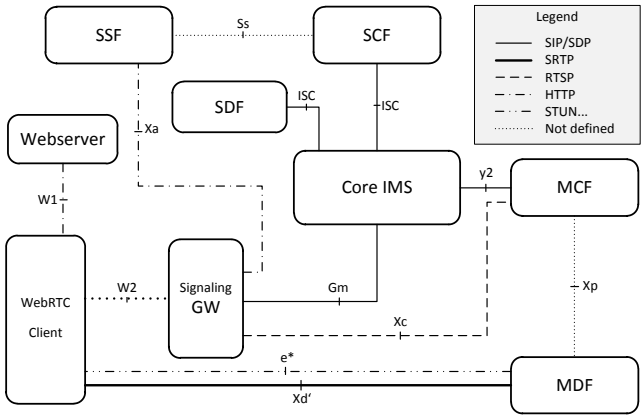


Figure 6: Architecture for WebRTC clients connected with an IMS-based IPTV ecosystem [1].

- Interfaces:
  - W1
  - W2
  - e*
  - Xp
  - Xd (modified to Xd')
  - Gm
  - Xa
  - Ss
  - ISC
  - y2

The web server is needed to provide the WebRTC application sources. The WebRTC application is executed in a WebRTC capable browser. The architecture of a WebRTC client is depicted in Figure 3. The application provides signaling functions for the communication with the core network via the inserted Signaling Gateway (SGW). Therefore, to make IMS-based IPTV services accessible to WebRTC clients, Generic IPTV Capabilities described in [18] are supported. As described in [1], the SGW implements some of these generic capabilities.

This gateway function converts session control messages coming from the WebRTC client side into Session Initiation Protocol (SIP) messages for the IMS core network side and vice versa. The SGW generates and forwards SIP messages towards the IMS core network and acts in place of the WebRTC client as a SIP capable signaling endpoint. As shown in Figure 6 the SGW also converts the session control messages from the WebRTC client into Hypertext Transfer Protocol (HTTP) and Real-Time Streaming Protocol (RTSP).

The core IMS is formed by the components, which are specified in [9]. This bulk of components in the core IMS implement several services, such as registration, provisioning, routing, accounting, billing, etc.. For more details of the components and their interworking and special function see [9].

The architecture includes also several IPTV functional components, which are standardized in [18]. The Service Discovery Function (SDF) in Figure 7 provides information about available services and related SSFs. The Service Selection Function (SSF) provides information that contains the metadata of the available content. The Service Control Function (SCF), is a SIP Application Server (AS) and the reference point for IMS UEs to start and control the IPTV sessions. The Media Control Function (MCF) controls the media transport of the MDF and receives instructions of the SCF and the UE. Also the selection of the right MDF is part of the MCF. The selection is made by several information, for example on codec information or geographical location. After a successful selection, the MCF transmits important session description information to the MDF. The Media Delivery Function (MDF) contains the media data and transmits them to the UE. In the architecture, the MDF is a modified MDF for WebRTC clients, which is proposed in [1]. It provides all necessary functions to establish a session with a WebRTC client. It supports STUN, ICE, DTLS and SRTP functionality. Also, a streaming engine with WebRTC capable codecs are implemented. These components are described more

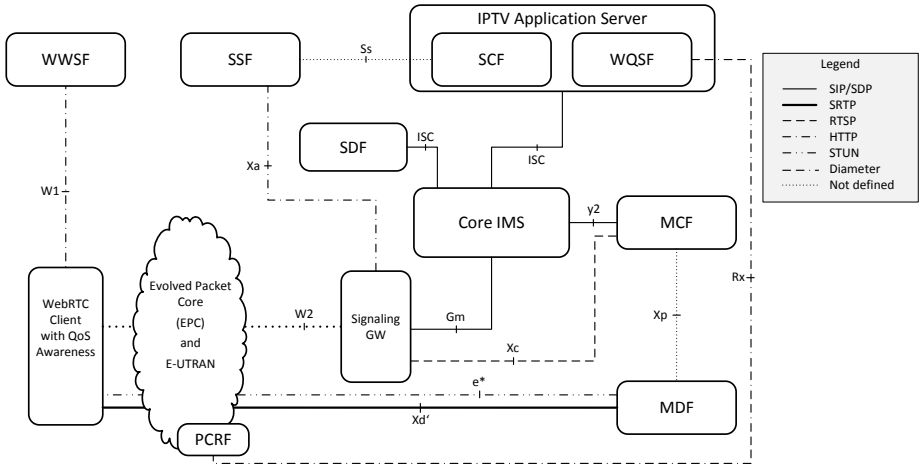


Figure 7: Consolidated architecture for IMS-based IPTV for QoS support.

in detail in [1].

The interfaces, shown in Figure 6, differ from the standardized architecture. The interfaces for the service interconnection are relocated from the user side to the SGW, which resides in the Telco's network infrastructure. The functionality and the used protocols of the Gm, Xc and Xa interfaces, located between the added SGW and the service or the core functions, are still conform to their specification [31]. The Ss, ISC and y2 interfaces remain unaffected. In addition to these interfaces above, some are modified or added and differ from the specification. These are described in more detail below.

The formerly undefined Xp interface between the MCF and the MDF is extended in the range of functions, respectively through the extension of the MDF.

The W1 interface is a reference point between the WebRTC client and the Webserver. It is used to download the application's source file using HTTP. Via this interface, the user receives the latest WebRTC web application.

The W2 interface is located between the WebRTC client and the SGW. The used protocol for this reference point is not defined and could be design by the developer. Therefore, the developer can choose from several state-of-the-art client-server protocols. The meaning of protocol messages regarding this interface must cover the sense of the transfered protocol messages from the interfaces Gm, Xa and Xc.

The Xd' interface between the WebRTC client and the MDF is responsible for media delivery using SRTP [32]. The original Xd interface only supports RTP/RTCP or HTTP for media delivery, so the modified interface for WebRTC interconnection is named Xd'. This modification results from the mandatory use of a secure connection in WebRTC [22].

The e* interface is a second reference point between the WebRTC client and the MDF. This interface is added in the concept of [1]. It is used for STUN connectivity checks between the WebRTC client side and the MDF. STUN is a mandatory to implement feature of WebRTC because the most WebRTC clients are behind NAT firewalls.

To enrich the whole environment with QoS, some components and interfaces have been added or modified. The resulting architecture is depicted in Figure 7 and its components and interface are described as follows:

- Components:
  - WWSF (added)
  - WebRTC client with QoS Awareness (modified)
  - IPTV Application Server combining SCF with WQSF (added)
  - EPC (added)
- Interfaces:
  - W2 (modified)
  - Rx (added)

In contrast to Figure 6, the WebRTC Web Server Function (WWSF) is depicted and represents the Webserver of the original architecture. The WWSF provides the actual WebRTC web application sources. This function may also be used for web-based authentication and authorization.

The WebRTC application is modified to support QoS. Therefore, to make QoS accessible to the WebRTC client, a special QoS awareness function is integrated. This integrated awareness function retrieves periodical and event-based flow information of the established and used media channels.

The IPTV Application Server acts as SIP AS combining the SCF with the WQSF, both functions are acting together as a single endpoint against the core IMS. The SCF part handles the IPTV specific messages and provides functions to control the session in a way which is described in [18]. The internal WQSF processes the WebRTC related flow information into Diameter-based QoS request messages and hence acts as an AF against the EPS network. More details of this WQSF are described in Figure 4 and in [2].

As access and control network we considered the EPC connected with the E-UTRAN. Together, they both build the overall EPS. To achieve QoS for the customers, the PCC concept, involving the PCRF and PCEF (as well as other internal components), is fully integrated into the EPS infrastructure.

The interface W2 is extended with specific QoS information messages. The functionality to support messages from the Gm, Xa and Xc remains unaffected.

A newly added interface is the Rx interface. This reference point is located between the IPTV AS, here the AF of the

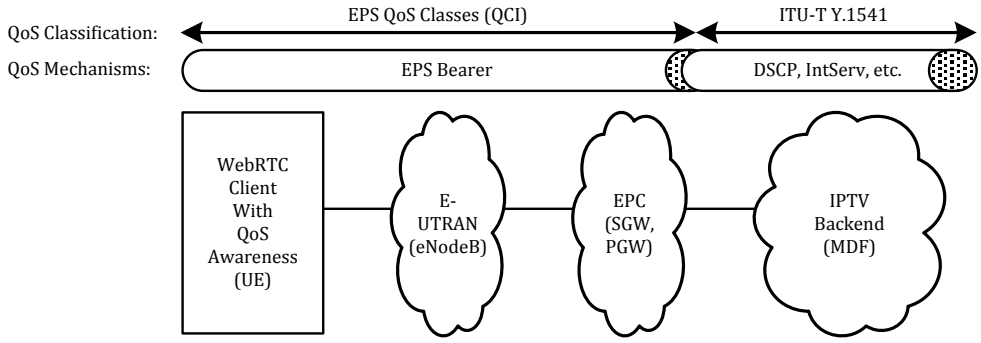


Figure 8: QoS Enforcement on the User Data Plane.

WQSF part of the AS, and a PCRF of the access network. The utilized protocol at this reference point is Diameter and it is used to transmit QoS related messages. The specific Diameter messages are described in [2] in more detail.

Figure 8 depicts the E2E chain for QoS enforcement between the UE and the MDF. In this chain two QoS classification schemes are applied, which exploits the mapping as proposed in Table IV. On the mobile network side, the user traffic treatment is categorized through the EPS QCI. To fulfill the QCI requirements, the user traffic is treated in an adequate EPS Bearer. On the IPTV Backend side, the user traffic treatment is categorized based on the recommendation of ITU-T Y.1541. For this, multiple QoS enforcement mechanisms can be used (e.g., DSCP, IntServ, MPLS). To achieve the expected quality on the consumer side, a mapping of the different QoS mechanisms is also necessary.

C. Proof of Concept (EPS and QoS with IMS IPTV over WebRTC)

Testbed components

To verify the functionality and the usability of the proposed concept, a testbed is prepared. With this testbed, the QoS initiation procedures for the content on demand use case (audio and video) is implemented and tested. However the actual enforcement of the QoS parameters on the user data plane is still outstanding.

For testing the concept the Google Chrome browser in version 38, which supports WebRTC, is used. The basis of this testbed is formed by an open-source IMS core network implementation originating from Fraunhofer FOKUS institute and now available on reference [33].

E-UTRAN functionality is provided by a LTE Femto Cell prototype with integrated eNodeB functionality by the company ip.access [34]. As mobile core network, an OpenEPC Rel. 3 testbed, initially developed from Fraunhofer FOKUS institute, is used [35].

The WWSF is an Apache HTTP Web Server, which provides the web application. The WebRTC client is implemented by using HTML5 and JavaScript. Based on this the Graphical User Interface (GUI) of the client is a responsive web site design using the jQuery mobile framework. This framework makes web sites accessible to all smart phone, tablet and desktop devices. The clients source code, based on JavaScript, utilizes the WebRTC API.

The SGW is written in C# and designed to handle several WebRTC Client sessions simultaneously. The prototyped SGW provides the main functionality for the interaction with the Gm and the Xa interface. The SGW firstly appears in [1] and is more sophisticated to cover the QoS support. Based on the sipsorcery project, an enhanced SIP protocol stack supporting IMS specific extensions is implemented [36].

Also, the IMS-based IPTV components are prototyped and inherited from [1]. All prototyped IPTV components, written in Java, are based on the technical specification [18]. The IPTV AS, containing the SCF and the WQSF, accepts the SIP requests via the JAIN-SIP stack. The SCF part handles all relevant SIP IPTV messages. The WQSF adapts the SIP requests containing the QoS relevant information into Diameter requests, based on JavaDiameterPeer library included in the OpenIMSCore project. The implementation of the MDF parses the session information, passed by the MCF. Within the MDF, the information are distributed to different engine functionalities. Open-source frameworks are used for

- The ICE agent with the STUN functionalities (icedjava) [37]
- The DTLS key exchange (BouncyCastle) [38],
- The SRTP implementation (srtplight) [39],
- And the streaming server (FFmpeg) [40].

Message Sequence

A sequence for Video on Demand over WebRTC with additional QoS reservation is depicted in Figure 9. The sequences can be described as follows:

Sequence 1) depicts the service discovery whereas the WebRTC client maps the service discovery messages into feasible messages to the Signaling Gateway via W2 and vice versa and whereas the Signaling Gateway maps the service discovery messages into feasible messages to the SDF via Gm and ISC and vice versa.

Sequence 2) depicts the service selection whereas the WebRTC client maps the service selection messages into feasible messages to the Signaling Gateway via W2 and vice

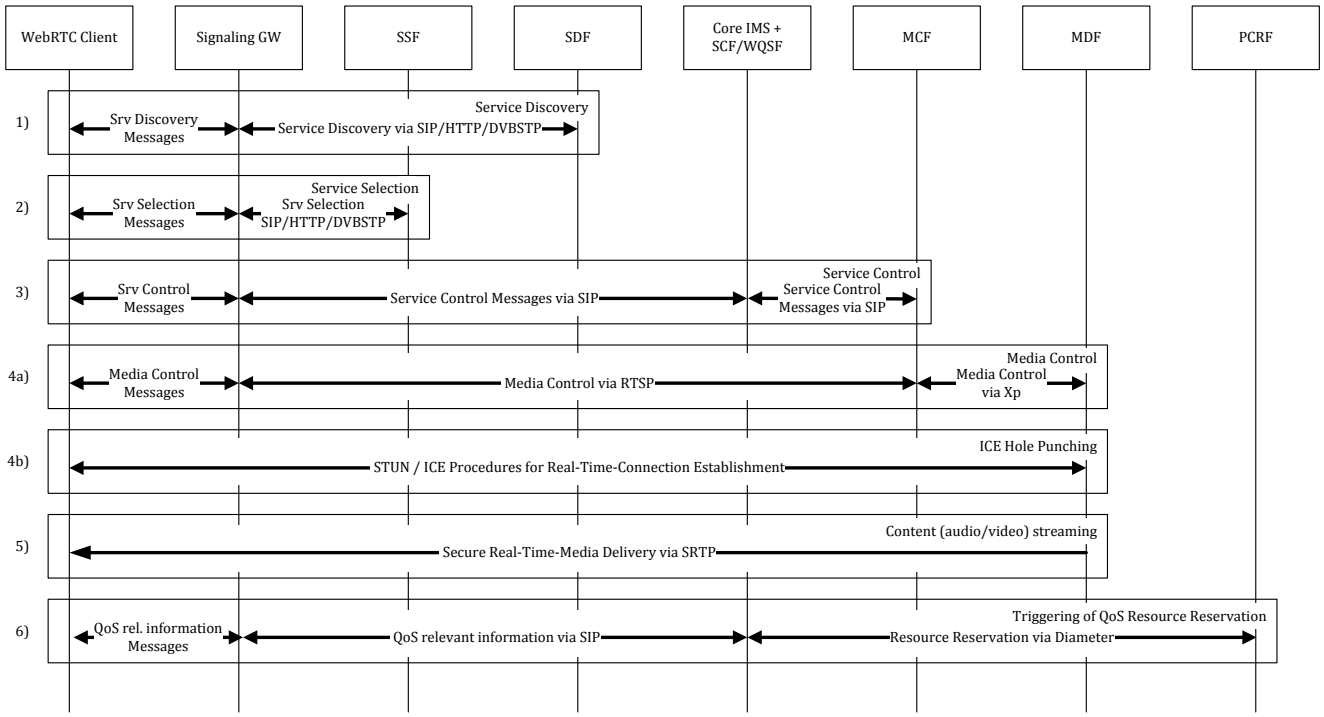


Figure 9: Sequence chart for WebRTC-based VoD with QoS resource reservation.

versa and whereas the Signaling Gateway maps the service selection messages into feasible messages to the SSF and via Xa vice versa. In the messages from the SSF, the client receives the service identifier to start the service.

Sequence 3) depicts the service control whereas the WebRTC client maps the service control messages into feasible messages containing the service identifier to the Signaling Gateway via W2 and vice versa and whereas the Signaling Gateway maps the service control messages into feasible messages to the IMS core via Gm and vice versa. With smart SIP routing the core transmits the messages to the SCF via IMS Service Control (ISC) by triggering on the service identifier. The SCF initiate the service delivery with session control messages via y2. Following session control message are transmitted from the user side to the SCF and vice versa.

Sequence 4a) depicts the media control whereas the WebRTC client maps the media control messages into feasible messages to the Signaling Gateway via W2 and vice versa and whereas the Signaling Gateway maps the media control messages into feasible messages to the MCF 16 via Xc and vice versa. Sequence 4a) further depicts the media control whereas the MCF maps the media control messages into feasible messages to the MDF via Xp and vice versa. The concurrent running sequence 4b) depicts the ICE/STUN procedures to establish a real-time path between the WebRTC Client and the MDF via e*.

Sequence 5) depicts the secured real-time streaming of audio/video between the WebRTC client and the MDF via Xd.

Sequence 6) depicts the QoS control process with signaling of QoS relevant information from the WebRTC client towards the targeted WQSF. The QoS relevant information is then used to request QoS related resources on the EPS network. This Sequence 6 is further described in the following subsection.

Triggering of QoS Resource Reservation

For triggering the QoS resource reservation, Figure 10 depicts a detailed message flow chart where the QoS relevant information is forwarded from the WebRTC client towards the QoS related network entities.

The UE's running WebRTC application retrieves all relevant information from the browser's internal *Browser RTC* function of the active WebRTC PeerConnection and extracts all connectivity pairs in which the user data is transmitted, received, or both. The application stores this connectivity information (including local and remote IP addresses as well as their transported media type) and prepares it into an overall JavaScript Object Notation (JSON) document, displayed in Listing 1. For each media flow, an array element will be added into the JSON document. As an overall QoS parameter, the document contains a field named *serviceType*, which incorporates the name of the used IPTV service. With that, all involved network entities can adapt the specific characteristics taken from Table IV for the IPTV application flow handling. In the example, the *serviceType* is set to *streaming_video_vod*. Analogue to this, it is also possible to use other IPTV services based on Table IV.

The presented JSON document (see Listing 1) will be relayed from the WebRTC client side into the SIP-based IMS environment. The Signaling Gateway enforces the conversion of the QoS relevant information received from the WebRTC proprietary signaling channel into standardized SIP messages (including the JSON document as SIP body).

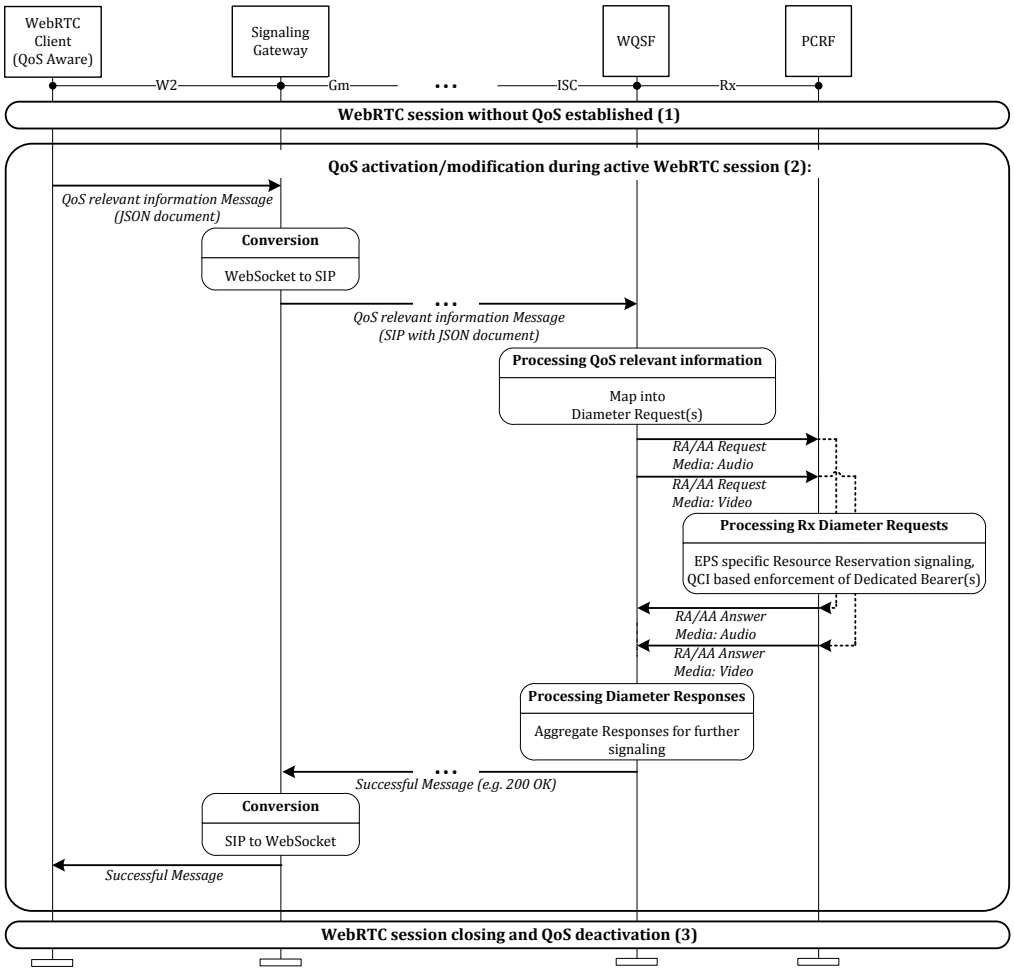


Figure 10: Sequence for QoS activation during a WebRTC session.

Listing 1: JSON document including QoS relevant information, provided by the WebRTC client.

```
{
 "cmd": "statisticInformationrequest",
 "attributes": {
  "sessionId": "406abf0d12174....",
  "localIdentity": "alice@domain.test",
  "messageType": "update",
  "remoteIdentity": "tvservice@domain.test",
  "serviceType": "streaming_video_vod"
  "flows": [
   {
   "remoteAddress": "192.168.7.4:34961",
   "localAddress": "192.168.13.63:56878",
   "mediaType": "audio"
   },
   {
   "remoteAddress": "192.168.7.4:40546",
   "localAddress": "192.168.13.63:56878",
   "mediaType": "video"
   }
  ]
 }
}
```

The WQSF performs the processing of the incoming JSON document. According to the content of the objects in the JSON document the WQSF generates related Diameter-based AVPs, as shown in Listing 2. For instance, the *AF-Application-Identifier* contains the intended *serviceType*: *streaming_video_vod*. Furthermore, the request also comprises the *Flow-Description* for the given video media flow.

Listing 2: Mapping of VoD with media type video into a Diameter Request.

```
Command Code: 265 AA-R (Request)
ApplicationId: 3GPP Rx (16777236)
...
AVP: Subscription-Id(443)
  AVP: Subscription-Id-Type(450)
    val=END_USER_SIP_URI (2)
  AVP: Subscription-Id-Data(444)
    val=sip:alice@domain.test
AVP: Media-Component-Description(517) vnd=TGPP
  AVP: Media-Type(520) vnd=TGPP val=VIDEO (1)
  AVP: Media-Sub-Component(519) vnd=TGPP
    AVP: Flow-Number(509) val=1
      AVP: Flow-Description(507) val=
        PERMIT OUT udp from 192.168.13.63 56878
```

```
        to 192.168.7.4 40546
      AVP: Flow-Description(507) val=
        PERMIT IN udp from 192.168.7.4 40546
        to 192.168.13.63 56878
...
AVP: AF-Application-Identifier(504) vnd=TGPP
  val=73747265616d696e675f766964656f5f766f64
  // Hex ASCII for: "streaming_video_vod"
```

Sequence 1 in Figure 10 shows the WebRTC session establishment phase. All procedures for this establishment such as the exchange of security keys, candidates for addressing issues etc. are covered in here but will not be described further. The first sequence is finished at this point. The establishment of the QoS characteristics will be processed in Sequence 2.

The second sequence depicts the actual QoS resource reservation interaction of the WebRTC session as well as the information sending process towards the WQSF. The WebRTC QoS Awareness Function processes QoS relevant session information of each media channel and forwards this as a JSON document towards the SGW. The SGW performs a conversion of the WebSocket message into a SIP request including the JSON document. This request will be forwarded through the core components towards the WQSF which maps this QoS relevant information into adequate Diameter requests. For each flow and media type (audio and video), described in the *flows*-array, the WQSF generates respective Re-Auth- or Authorise-Authenticate-*Request* (RAR/AAR) messages based on the JSON document [41]. The PCRF receives the Diameter messages and executes an resource reservation signaling based on the EPS QoS Class principles [13]. The proposed mapping is based on Table IV. Based on the given *AF-Application-Identifier* and the described *Media-Type* (audio and video), the QCI-based establishment and enforcement with Dedicated Bearers towards the PCEF inside the EPS core network. After processing the incoming Diameter Requests, the PCRF responses with Diameter-Answers towards the WQSF respectively. Inside the WQSF, the answers are stored and aggregated, related to that base JSON document. If all requests are answered successfully, the WQSF replies with a *successful message* (i.e., SIP Response 200 OK) along the initial way through the core network components. Finally, the SGW performs a conversion of the SIP response into WebSocket message.

The third sequence depicts the session closing procedures and has an analogue behavior. Initiated by the WebRTC client side, the session closing and QoS deactivation process follows same pattern as Sequence 2.

IV. Conclusion

The authors showed how QoS management could work for a WebRTC client which is connected to a mobile 4G network. In particular, we proposed a novel concept, which includes the following:

- An adequate QoS class mapping (see also Table IV), which can be implemented into all involved network elements (4G mobile and IPTV core network);
- An aggregated network architecture (see also Figure 7);
- A detailed QoS control concept, which is already prototyped and tested in practice (see also Figure 10);
- A concept for E2E QoS enforcement on the user data plane.

Considering an use case for live TV, respectively VoD, it is pointed out how WebRTC clients could benefit from a resource reservation rather than having no QoS allocation. For IPTV services like Near VoD (e.g., YouTube watching) a QoS support as proposed in Table IV should be helpful for mobile 4G end-users. We see a high potential to combine our WebRTC-based QoS concept with the VoD services utilizing recent HTML5 technologies. The technical details for a successful consolidation are for further study.

The following aspects are relevant within this journal contribution and are still outstanding at the moment:

- Implementation of the QoS relevant enforcement components in the User Plane and their performance testing;
- Analyzing the subjective end user expectations correlating to the different IPTV services;
- Involving other IPTV quality assessment methods (instrumental and perceptual).

The authors are positive about the increasing relevance of end-to-end QoS in heterogeneous networks in the near future.

Furthermore, several fifth generation (5G) white-papers forecast the heavily growing need for QoS resources for the next years [42]. With the identified new upcoming technologies therein, such as 3D audio, 3D video and ultra-high-definition formats and codecs, the demand for lower latencies and higher per-user data rates will increase tremendously. In that process, it is important to detect the necessary QoS demands for each contexts individually. Not until then, it is possible to deal fairly on the finite network resources and serve all users with the best experience for their individual real-time applications.

QoS should be made accessible, not only statical by network providers pre-configured parameters but also dynamically allocatable through the users applications regarding his current state of communication. The author's proposed concept, providing an interface for the user applications to request for a QoS reservation, could be a solution for the future.

Acknowledgment

The Telekom Innovation Laboratories (T-Labs) and the Hochschule fuer Telekommunikation Leipzig (HfTL) are actively cooperating since 2011. Both are working together on common topics in the area of fixed-/mobile converged network infrastructure including IMS-based services. They are participating in ongoing projects relating WebRTC with Telco Assets like QoS and interoperability. This paper also presents some of the results and acquired experience arising from [1] and [2].

References

[1] T. Bach, M. Maruschke, J. Zimmermann, K. Haensge, and M. Baumgart, "Combination of IMS-based IPTV services with WebRTC," ICCGI 2014, The 9th International Multi-Conference on Computing in the Global Information Technology, IARIA, June 2014, pp. 140–145.

[2] K. Haensge and M. Maruschke, "QoS-based WebRTC access to an EPS network infrastructure," in 2015 18th International Conference on Intelligence in Next Generation Networks (ICIN 2015), Paris, France, Feb. 2015, pp. 9 – 15.

[3] C.Aras, J.Kurose, D.Reeves, and H.Schulzrinne., "Real-time communication in packet-switched networks," Proceedings of the IEEE, Jan 1994, pp. 122 – 139.

[4] ITU-T, "Internet protocol aspects Quality of service and network performance;Internet protocol data communication service IP packet transfer and availability performance parameters," International Telecommunication Union (Telecommunication Standardization Sector),

REC Y.1540, Nov. 2007. [Online]. Available: http://www.itu.int/rec/T-REC-Y.1540-200711-S/en

[5] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 3550 (Standard), Internet Engineering Task Force, Jul. 2003, updated by RFCs 5506, 5761, 6051, 6222. [Online]. Available: http://www.ietf.org/rfc/rfc3550.txt

[6] ITU-T, "TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS, International telephone connections and circuits General Recommendations on the transmission quality for an entire international telephone connection-One-way transmission time," International Telecommunication Union (Telecommunication Standardization Sector), REC G.114, May 2003. [Online]. Available: http://www.itu.int/rec/T-REC-G.114-200305-I/en

[7] ——, "Quality of telecommunication services: concepts, models, objectives and dependability planning Terms and definitions related to the quality of telecommunication services; Definitions of terms related to quality of service," International Telecommunication Union (Telecommunication Standardization Sector), REC E.800, Sep. 2008. [Online]. Available: http://www.itu.int/rec/T-REC-E.800-200809-I/en

[8] ——, "Internet protocol aspects Quality of service and network performance;Network performance objectives for IP-based services," International Telecommunication Union (Telecommunication Standardization Sector), REC Y.1541, Dec. 2011. [Online]. Available: http://www.itu.int/rec/T-REC-Y.1541-201112-I/en

[9] 3GPP, "IP Multimedia Subsystem (IMS); Stage 2," 3rd Generation Partnership Project (3GPP), TS 23.228 v11.4.0, Mar. 2012. [Online]. Available: http://ftp.3gpp.org/specs/html-info/23228.htm

[10] ITU-T, "Next Generation Networks Frameworks and functional architecture models;General overview of NGN," International Telecommunication Union (Telecommunication Standardization Sector), REC Y.2001, Dec. 2004. [Online]. Available: http://www.itu.int/rec/T-REC-Y.2001/en

[11] 3GPP, "Quality of Service (QoS) concept and architecture," 3rd Generation Partnership Project (3GPP)(Release8), TS 23.107 v8.2.0, Sep. 2011. [Online]. Available: http://www.3gpp.org/DynaReport/23107.htm

[12] ——, "E-UTRA and E-UTRAN overall description," 3rd Generation Partnership Project (3GPP), TS 36.300 v8.12.0, Mar. 2010. [Online]. Available: http://ftp.3gpp.org/Specs/html-info/36300.htm

[13] ——, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 8)," 3rd Generation Partnership Project (3GPP), TS 23.203 v8.14.0, Jun. 2012. [Online]. Available: http://www.3gpp.org/DynaReport/23203.htm

[14] ETSI, "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Resource and Admission Control Sub-System (RACS): Functional Architecture," European Telecommunications Standards Institute (ETSI), ES 282003 v3.5.1, Apr. 2011. [Online]. Available: http://pda.etsi.org/pda/queryform.asp

[15] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," RFC 2474 (Proposed Standard), Internet Engineering Task Force, Dec. 1998, updated by RFCs 3168, 3260. [Online]. Available: http://www.ietf.org/rfc/rfc2474.txt

[16] ETSI, "Speech and multimedia Transmission Quality (STQ); QoS and network performance metrics and measurement methods; Part4," European Telecommunications Standards Institute (ETSI), ES 202765-4 v1.2.1, May 2014. [Online]. Available: http://pda.etsi.org/pda/queryform.asp

[17] ITU-T, "Objective perceptual multimedia video quality measurement in the presence of a full reference," International Telecommunication Union (Telecommunication Standardization Sector), REC J.247, Aug. 2008. [Online]. Available: http://www.itu.int/rec/T-REC-J.247-200808-I/en

[18] ETSI, "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IPTV Architecture; IPTV functions supported by the IMS subsystem," European Telecommunications Standards Institute (ETSI), TS 182027 v3.5.1, Mar. 2011, Available: http://www.etsi.org/deliver/etsi_ts/182000_182099/182027/03.05.01_60/ts_182027v030501p.pdf [retrieved: May, 2015].

[19] ITU-T, "IPTV over NGN Guidelines for the use of traffic management mechanisms in support of IPTV services," International Telecommunication Union (Telecommunication Standardization Sector), REC Y.1920, Jul. 2012. [Online]. Available: http://www.itu.int/rec/T-REC-Y.1920-201207-I

[20] C. Jennings, A. Narayanan, D. Burnett, and A. Bergkvist, "WebRTC 1.0: Real-time communication between browsers," W3C, W3C Working Draft, Feb. 2015, http://www.w3.org/TR/2015/WD-webrtc-20150210/.

[21] IETF. Rtcweb status pages. [Online]. Available: http://tools.ietf.org/wg/rtcweb/ [retrieved: May, 2015]

[22] C. Perkins, M. Westerlund, and J. Ott, "Web Real-Time Communication (WebRTC): Media Transport and Use of RTP draft-ietf-rtcweb-rtp-usage-11," Internet-Draft, Internet Engineering Task Force, Dec. 2013, Available: http://tools.ietf.org/id/draft-ietf-rtcweb-rtp-usage-23.txt [retrieved: May, 2015].

[23] E. Rescorla and N. Modadugu, "Datagram Transport Layer Security Version 1.2," RFC 6347 (Proposed Standard), Internet Engineering Task Force, Jan. 2012, Available: http://www.ietf.org/rfc/rfc6347.txt [retrieved: May, 2015].

[24] E. Rescorla, "WebRTC Security Architecture draft-ietf-rtcweb-security-arch-07," Internet-Draft, Internet Engineering Task Force, Jul. 2013, Available: http://tools.ietf.org/id/draft-ietf-rtcweb-security-arch-11.txt [retrieved: May, 2015].

[25] J. Rosenberg, R. Mahy, P. Matthews, and D. Wing, "Session Traversal Utilities for NAT (STUN)," RFC 5389 (Proposed Standard), Internet Engineering Task Force, Oct. 2008, Available: http://www.ietf.org/rfc/rfc5389.txt [retrieved: May, 2015].

[26] J. Rosenberg, "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols," RFC 5245 (Proposed Standard), Internet Engineering Task Force, Apr. 2010, updated by RFC 6336. Available: http://www.ietf.org/rfc/rfc5245.txt [retrieved: May, 2015].

[27] H. Alvestrand, "Overview: Real Time Protocols for Brower-based Applications draft-ietf-rtcweb-overview-08," Internet-Draft, Internet Engineering Task Force, Sep. 2013, Available: http://tools.ietf.org/id/draft-ietf-rtcweb-overview-13.txt [retrieved: May, 2015].

[28] "Web Real-Time Communication Use Cases and Requirements," RFC 7478 (Informational), Internet Engineering Task Force, Mar. 2015. [Online]. Available: http://www.ietf.org/rfc/rfc7478.txt

[29] S. Dhesikan, C. Jennings, D. Druta, P. Jones, and J. Polk, "DSCP and other packet markings for WebRTC QoS: draft-ietf-tsvwg-rtcweb-qos-03," Internet-Draft, Internet Engineering Task Force, Nov. 2014, Available: https://tools.ietf.org/html/draft-ietf-tsvwg-rtcweb-qos-03 [retrieved: May, 2015].

[30] International Organization for Standardization, "Information technology - Open Systems Interconnection - Basic Reference Model," International Organization for Standardization, REC ISO/IEC 7498-1:1994, Nov. 1994. [Online]. Available: http://www.iso.org

[31] ETSI, "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS-based IPTV stage 3 specification," European Telecommunications Standards Institute (ETSI), TS 183063 v3.5.2, Mar. 2011, Available: http://www.etsi.org/deliver/etsi_ts/183000_183099/183063/03.05.02_60/ts_183063v030502p.pdf [retrieved: May, 2015].

[32] M. Baugher, D. McGrew, M. Naslund, E. Carrara, and K. Norrman, "The Secure Real-time Transport Protocol (SRTP)," RFC 3711 (Proposed Standard), Internet Engineering Task Force, Mar. 2004, updated by RFC 5506. Available: http://www.ietf.org/rfc/rfc3711.txt [retrieved: May, 2015].

[33] Openimscore.org. Opensourceims. [Online]. Available: http://www.openimscore.org/ [retrieved: May, 2015]

[34] ip.access Ltd. ip.access: leaders in 2g, 3g and 4g end-to-end small cell solutions. [Online]. Available: http://www.ipaccess.com/ [retrieved: May, 2015]

[35] OpenEPC. Openepc.com. [Online]. Available: http://www.openepc.com/ [retrieved: May, 2015]

[36] T. Bach. sipsorcery-fork. [Online]. Available: https://github.com/hftl-ims-research/sipsorcery-fork [retrieved: May, 2015]

[37] inspired social. Open source ICE implementation. [Online]. Available: http://code.google.com/p/inspired-social/source/browse/trunk/StunServer/net/mc_cubed/icedjava?spec=svn20&r=20#icedjava [retrieved: May, 2015]

[38] Legion of the Bouncy Castle Inc. Bouncy Castlee Crypto API. [Online]. Available: http://www.bouncycastle.org [retrieved: May, 2015]

[39] steely gint. set of classes implementing a simple (S)RTP stack. [Online]. Available: https://github.com/steely-glint/srtplight [retrieved: May, 2015]

[40] FFmpeg: a open source cross-platform solution to record, convert and stream audio and video. [Online]. Available: http://www.ffmpeg.org/index.html [retrieved: May, 2015]

[41] "Diameter Network Access Server Application," RFC 7155 (Proposed Standard), Internet Engineering Task Force, Apr. 2014. [Online]. Available: http://www.ietf.org/rfc/rfc7155.txt

[42] 4GAmericas. 4G Americas' Recommendations on 5g Requirements and Solutions. [Online]. Available: http://www.4gamericas.org/en/resources/white-papers/ [retrieved: May, 2015]

On Type II Hybrid-ARQ with Decode and Forward Relay using Non-Binary Rate-Compatible Punctured LDPC Code on MIMO Frequency Selective Channels

Sho Kato and Yasunori Iwanami
Dept. of Computer Science and Engineering
Nagoya Institute of Technology
Nagoya, Japan
E-mail: 26417542@stn.nitech.ac.jp, iwanami@nitech.ac.jp

***Abstract*— In this paper, Non-Binary Rate-Compatible Punctured Low Density Parity Check (NB RCP LDPC) code is designed over the extended Galois Field. The designed NB RCP LDPC code is applied to the type II Hybrid Automatic Repeat reQuest (HARQ) with Decode and Forward (DF) relay on Multiple Input Multiple Output (MIMO) Orthogonal Frequency Division Multiplexing (OFDM) channel and MIMO Single Carrier-Frequency Division Multiple Access (SC-FDMA) channel. The designed code enables us to decrease the coding rate with incremental redundancy for each retransmission in HARQ. The retransmission is done from the DF relay after the successful decoding in the relay. We have verified through computer simulations that the proposed type II HARQ scheme with DF relay greatly improves the throughput and average retransmission characteristics compared with the scheme without DF relay. Multiple relay cases are also considered.**

Keywords-NB RCP LDPC code; Hybrid-ARQ; Decode and Forward Relay; MIMO; OFDM; SC-FDMA; Symbol-LLR.

I. INTRODUCTION

An LDPC code that suits the flexible coding rate design and has the high error correcting capability through iterative decoding can be constructed on arbitrary extended Galois Field (GF). The Non-Binary (NB) LDPC code constructed on extended GF [1],[2] generally exhibits the better Bit Error Rate (BER) performance than the binary LDPC codes [3],[4]. There also exist RCP LDPC codes with variable coding rate obtained by properly puncturing the mother LDPC code [5]. The RCP LDPC codes enable us to use the same decoder as the mother code and suit the ARQ error correcting schemes [6],[7] with the incremental redundancy. By combining the NB LDPC codes with the RCP codes, the NB RCP LDPC codes were designed and the designed NB RCP LDPC codes were applied to the type II HARQ [8],[9]. When comparing the HARQ using NB RCP LDPC codes with the existing RCPT (Rate Compatible Punctured Turbo) HARQ using binary Turbo codes [10], the HARQ with NB RCP LDPC codes can cope with flexible coding rates, code word lengths and NB symbol LLR [11] additions without using inter-leavers for burst errors on the channel. On the other hand, the Decode and Forward (DF) relay schemes [12],[13] are useful for HARQ schemes. By using the DF relay, the source node can be replaced by the relay, once the relay correctly decodes the packet from the source. This replacement from the source to the relay effectively reduces the number of retransmissions and improves the throughput. In [9], NB LDPC coding with NB repletion codes is applied to multiple relay case for flat fading channel. The NB RCP LDPC coded type II HARQ with DF relay is applied to the MIMO-OFDM modulation in [2]. The incremental redundancy in HARQ with DF relay is also suited to the up-link transmission like in Long Term Evolution (LTE) or 4G. Due to the necessities of low PAPR and the high power efficiency in the amplification, MIMO SC-FDMA [14] is usually adopted to the up-links in cellular networks. Among SC-FDMA, interleaved SC-FDMA is especially useful because of its very low PAPR nature and excellent frequency diversity effect [14]. The application of NB RCP LDPC codes to MIMO interleaved SC-FDMA with multiple DF relays was reported in [1].

In this paper, we have examined in detail the performance of NB RCP LDPC coded type II HARQ with DF relays on MIMO OFDM channel and MIMO interleaved SC-FDMA channel with completely new simulations. Although some parts in this paper are identical to [1] and [2], the contents of [1] and [2] are verified and some new simulations and results are added. We have confirmed that the proposed HARQ scheme with DF relay greatly improves the throughput and the average number of retransmission characteristics compared with the case of no DF relay. Moreover, we have investigated the multiple relay cases.

The paper is organized as follows. In Section II, RCP LDPC code is introduced. In Section III, NB LDPC coded Type II HARQ scheme is described. In Section IV, we introduce the DF relaying scheme. In Section V, we briefly illustrate the MIMO OFDM and MIMO interleaved SC-FDMA modulation. In Section VI, computer simulation results are shown. The paper concludes with Section VII.

II. RCP LDPC CODE

The encoding and decoding procedure of RCP LDPC code is as follows. We call the code before puncture and the code after puncture as the mother code and the efficient code, respectively. In RCP LDPC code, the encoder and decoder of mother code can also be applied to the efficient code. When the parity check matrix of mother code is given by $\boldsymbol{H}_M(n_M \times n_N)$ and the generator matrix by $\boldsymbol{G}_M(n_N \times n_K)$ with $n_K = (n_N - n_M)$, the coding rate of mother code becomes $R_M = (1 - n_M / n_N) = n_K / n_N$. The coding rate after the puncture of n_P symbols from the mother code is given

by $R_E = n_K / (n_N - n_P)$. We denote the message vector as $\boldsymbol{m} = (m_1, m_2, \ldots, m_{n_K})$, the code word of mother code as $\boldsymbol{C_M} = (c_{M1}, c_{M2}, \ldots, c_{Mn_N})$, the index of position to be punctured as $\boldsymbol{P} = (p_1, p_2, \ldots, p_{n_P})$ and the code word of punctured code as $\boldsymbol{C_P} = (c_{P1}, c_{P2}, \cdots, c_{Pn_P})$. The encoding procedure is first to generate the mother code by $\boldsymbol{C_M} = \boldsymbol{mG_M}$ which is systematic, and next, to puncture the position using $\boldsymbol{P}$ to obtain $\boldsymbol{C_P}$. The decoding procedure is to produce the symbol LLR (Log Likelihood Ratio) [11] from the receive signal and it is fed to the mother code decoder as the initial value for the sum-product algorithm. The symbol LLR for the position $\boldsymbol{P}$ is initially set to 0, because there is no available symbol LLR corresponding to the position $\boldsymbol{P}$.

III. NB RCP LDPC CODED TYPE II HARQ SCHEME

In Figure 1, we show the transmitter and receiver block diagram of NB RCP LDPC coded Type II HARQ. At the transmitter, the data bits are firstly encoded by the Cyclic Redundancy Check (CRC)-16 error detecting code and secondly encoded by the NB LDPC code on GF(4) or GF(16). The encoded LDPC code word is divided into the transmission packets and they are modulated by Quaternary Phase Shift Keying (QPSK) or 16 Quadrature Amplitude Modulation (16QAM) depending on GF(4) or GF(16), respectively. Matching GF(Q) to the modulation level Q is preferable in calculating the symbol LLR and reduces the complexity compared with the use of bit LLR calculation. After the interpolation filtering and the up-conversion to carrier frequency, the Radio Frequency (RF) signal is transmitted from the antenna.

At the receiver side, the received signal is demodulated and the symbol LLR is calculated. The symbol LLR is then fed to the NB LDPC decoder. Using Sum-Product Algorithm (SPA), the LDPC code word is decoded and the hard decision is made to obtain the data bits. The data bits

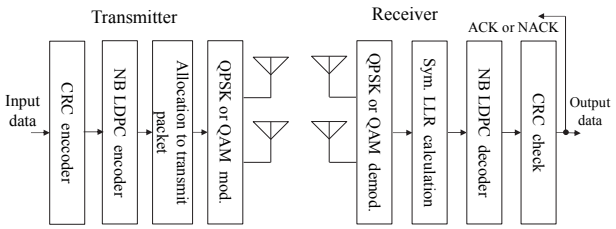


Figure 1. Transmitter and receiver structure of NB RCP LDPC coded type II HARQ scheme

TABLE I. SIMULATION CONDITIONS OF RCP LDPC CODE

Channel		AWGN	
Modulation		QPSK	16QAM
Size of Galois field		GF(4)	GF(16)
Mother code	Size of parity check matrix	(256,512)	(128,256)
	Average weight	(2.66,5.32)	(2.41,4.82)
	Coding rate	4/8	2/4
Efficient code	Information bit length	512	
	Coding rate	4/8,4/7,4/6,4/5, 4/4	2/4,2/3,2/2
Max SPA iteration		20	

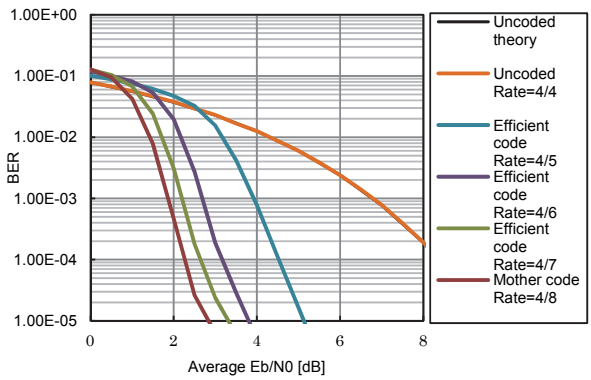


Figure 2. BER characteristics of NB RCP-LDPC code on AWGN channel (QPSK, Uncoded theory and uncoded rate 4/4 are overlapped.)

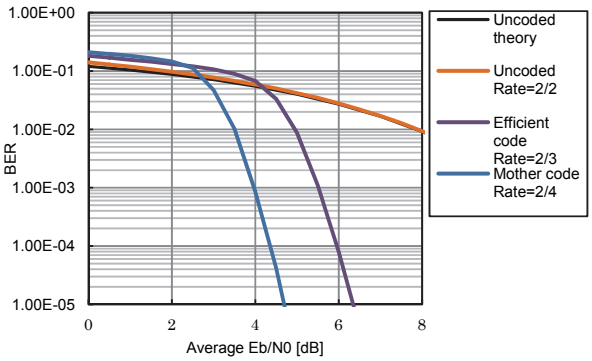


Figure 3. BER characteristics of NB RCP-LDPC code on AWGN channel (16QAM, Uncoded theory and uncoded rate 2/2 are overlapped.)

are then CRC-checked and Negative ACKnowledgement (NACK) or ACKnowledgement (ACK) is returned to the transmitter corresponding to the error or no error detection. The BER characteristics of RCP LDPC code on Additive White Gaussian Noise (AWGN) channel are examined when the rate 1/2 mother code on GF(4) or GF(16) is punctured to change the coding rate. The simulation condition is listed in Table I and the simulation results are shown in Figure 2 and Figure 3. C++ language is utilized for programming. From the simulation results, we know that the efficient codes on GF(4) or GF(16) with different coding rates are obtained from a mother code and the error correction capability corresponding to each coding rate is achieved.

In type II HARQ, like in Figure 4, at the first transmission, only uncoded information symbols are transmitted, and at the second transmission and after, the parity check symbols are retransmitted at the incremental redundancy policy. Accordingly, when the channel condition is good, the first uncoded transmission is successful and it achieves high throughput. On the other hand, when the channel condition is bad, by decreasing the coding rate at each retransmission, the error correction capability increases gradually. The generation of NB RCP LDPC code is done only once at the transmitter and there is no need of regeneration of code word when the coding rate

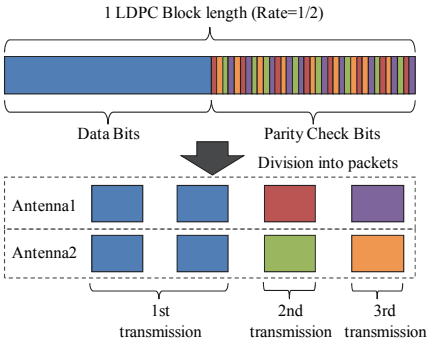


Figure 4. Division of NB LDPC code word into transmission packets when 2 spatial multiplex streams are employed

is decreased. Therefore, there is no increase of complexity of NB RCP LDPC code compared with the fixed rate NB LDPC code. Also at the receiver side, the complexity of NB RCP LDPC decoder does not increase compared with the fixed rate NB LDPC decoder, because the same and only one NB LDPC decoder can be used for various coding rates of NB RCP LDPC code.

IV. DECODE AND FORWARD RELAYING SCHEME

The Decode and Forward relay model is shown in Figure 5. We consider the relay arrangement where the relay locates at the middle point between the source (transmitter) and the destination (receiver).

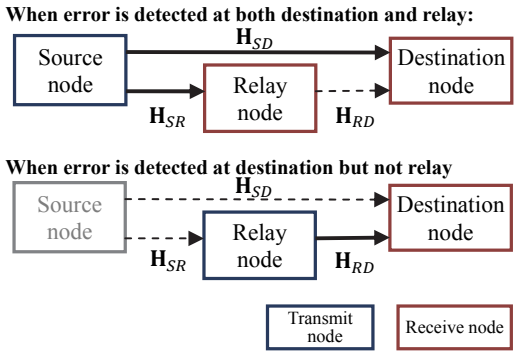


Figure 5. DF (Decode and Forward) relaying model

At the first transmission, the source broadcasts the uncoded information packet to the relay and destination simultaneously. The relay and destination independently detect the transmission errors using CRC-16 code. The relay and destination independently return (broadcast) ACK or NACK to the source. This ACK or NACK is shared among source, relay and destination. If the destination returns ACK, the transmission finishes at the first transmission and this condition is equivalent to no relay. Otherwise, retransmission is made. The source sends parity check packets with incremental redundancy. The relay and destination receive the parity check packet and combine it with already received packet. The LDPC decoding and CRC error detection is done both at relay and destination. ACK or NACK is returned and shared among source, relay and destination. At this point, if destination returns NACK but relay does ACK, then the relay sends the parity check packet hereafter instead of the source, i.e., the source is replaced by the relay which locates closer to destination. The transmission from relay to destination is more successful than source to destination due to the near distance between relay and destination. Also, as the source and relay do not simultaneously retransmit the parity check packet, the total transmission power is the same between with and without relay. This saves the total transmit energy in the case when the same transmit power as source is allocated to the relay.

Next, we consider the two relay cases where two relays are allocated in parallel or serial manner as shown in Figure 6(b) or (c), respectively. Figure 6 (a) is the arrangement of single relay already discussed. In Figure 6 (b), two relays are allocated in the middle point between source and destination in parallel. In Figure 6(c), relay 1 and relay 2 are allocated serially with equal distance interval between source and destination. When the power attenuation exponent is given by α and the distance between source and destination is normalized as 1, the relay at the middle point between source and destination in Figure 6 (a) and (b) receives 2^{α} times more power than the direct link between source and destination. Similarly, the relay 1 and relay 2 in Figure 6 (c) receives 3^{α} times more power than the direct link.

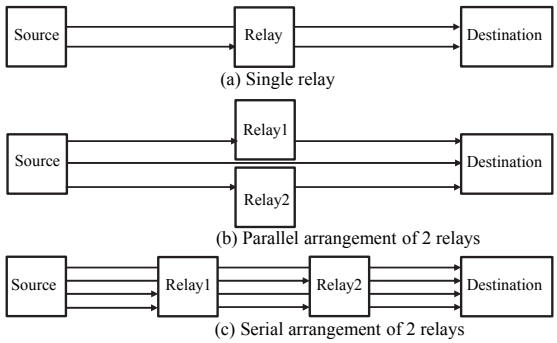


Figure 6. DF relay arrangement in case of multiple relays

V. MIMO OFDM AND MIMO SC-FDMA MODULATION SHEME

In Figure 7, we show the block diagram of NB RCP LDPC coded Type II HARQ scheme using MIMO-OFDM modulation. At the transmitter, the data bits are firstly encoded by the CRC-16 error detecting code and secondly encoded by the NB LDPC code on GF(4) or GF(16). The encoded NB alphabets are mapped to QPSK signal points for GF(4) or 16QAM for GF(16). These signal points are then modulated by OFDM with guard interval insertion for making a packet for each transmission. The OFDM signal is then transmitted to the frequency selective channel from each transmit antenna. At the receiver, for each antenna, the guard interval is first removed and then OFDM demodulation is made using Fast Fourier Transform (FFT). By demodulating each subcarrier QPSK modulated or

16QAM modulated, the symbol LLR is calculated.

We show how the symbol LLR is calculated for each demodulated subcarrier of OFDM with QPSK signaling. When the transmit signal, receive signal, signal points of QPSK and the subcarrier channel fading value are denoted as x, r, s_0, s_1, s_2, s_3 and h, respectively, the symbol LLR for the alphabets $a = 0,1,2,3$ on GF(4) is defined as

$$LLR_a = \log_e\left\{\frac{P(x=a|r\Delta r)}{P(x=0|r\Delta r)}\right\} = \log_e\left\{\frac{P(s_a, r\Delta r)/P(r\Delta r)}{P(s_0, r\Delta r)/P(r\Delta r)}\right\}$$
$$= \log_e\left\{\frac{P(s_a, r\Delta r)}{P(s_0, r\Delta r)}\right\} = \log_e\left\{\frac{P(s_a)p(r\Delta r|s_a)}{P(s_0)P(r\Delta r|s_0)}\right\} \quad (1)$$
$$= \log_e\left\{\frac{p(r|s_a)\Delta r}{P(r|s_0)\Delta r}\right\} = \log_e\frac{p(r|s_a)}{P(r|s_0)}$$

where the priori probabilities are set to $P(s_0) = P(s_1) = P(s_2) = P(s_3) = 1/4$, i.e., equal probabilities. In (1), $P(x=a|r\Delta r)$ denotes the probability that the transmit symbol x equals a when the receive signal point r falls in the small area $r\Delta r$ centered at r. $p(r|s_a)$ is the transition probability density function from $s_a \to r$ and is expressed as

$$p(r|s_a) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|r-hs_a|^2}{2\sigma^2}\right) \quad (2)$$

where σ^2 is the variance of receive noise. Accordingly, the symbol LLR is calculated as

$$LLR_a = \log_e\left\{\frac{p(r|s_a)}{p(r|s_0)}\right\} = \frac{|r-hs_0|^2 - |r-hs_a|^2}{2\sigma^2} \quad (3)$$

The symbol LLR values are then fed to the LDPC decoder and the iterative decoding using sum-product algorithm is done. The decoded information bits are error-detected by the CRC-16 code. If error is not detected, the data bits are fed to the data sink and the ACK is returned to the transmitter to finish the transmission. But if errors are detected, the NACK is returned and the retransmission is requested. As the type II HARQ scheme is employed, at the first transmission, only the data symbols without encoding are sent to the receiver. After the 2nd transmission, as shown in Figure 4, the parity symbols are sent several times with the incremental redundancy depending on the error detection status at the receiver. When the channel quality is good, the uncoded data packet for the first transmission succeeds with high probability leading to the high throughput performance. On the other hand, when the channel quality is bad, the parity packets are retransmitted several times till the LDPC code rate reaches the lowest one

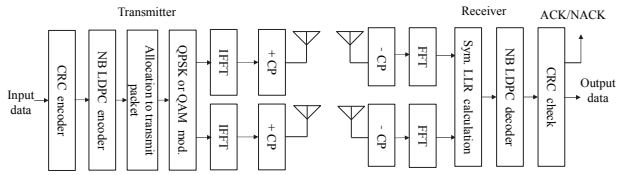


Figure 7. Transmitter and receiver configuration of NB RCP LDPC coded type II HARQ scheme using MIMO OFDM

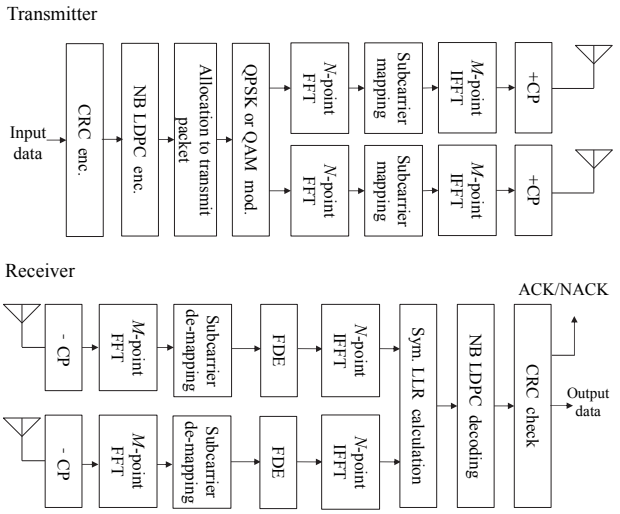


Figure 8. Transmitter and receiver configuration of NB RCP LDPC coded type II HARQ scheme using MIMO interleaved SC-FDMA

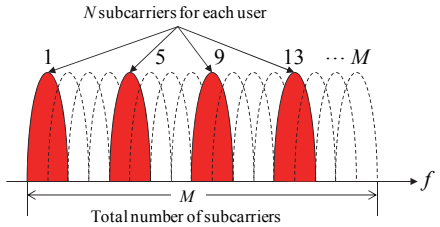


Figure 9. Subcarrier mapping in interleaved SC-FDMA

half resulting in enough error correction capability.

In Figure 8, we show the transmitter and receiver block diagram of NB RCP LDPC coded Type II HARQ using MIMO SC-FDMA. At the transmitter, the data bits are firstly encoded by the Cyclic Redundancy Check (CRC)-16 error detecting code and secondly encoded by the NB LDPC code on GF(4) or GF(16). The encoded LDPC code word is divided into the transmission packets and they are modulated by QPSK or 16QAM depending on GF(4) or GF(16), respectively. The modulated QAM symbols are then N-point FFT transformed at each antenna stream and the subcarrier mapping is done to make the interleaved SC-FDMA spectrum as shown in Figure 9. The interleaved spectrum is then M-point Inverse FFT (IFFT) transformed, where $M = U \times N$ and U is the number of users ($U = 4$ in Figure 9). Cyclic Prefix (CP) is added to the time domain complex samples of an IFFT block. After the interpolation filtering and the up-conversion to carrier frequency, the RF signal is transmitted from each antenna. At the base station, after the down-conversion to baseband and the sampling, CP is removed and the M-point FFT is done to obtain the frequency domain signal. The frequency domain signal is then subcarrier-de-mapped to aggregate the interleaved spectrum of each user back to the N sample spectrum again. The Frequency Domain Equalization (FDE) is made to compensate the channel frequency response and separate

the multiple spatial streams of each user. After the FDE, *N*-point IFFT is made to obtain the time domain signal of each stream. The subsequent symbol LLR calculation, NB LDPC decoding, CRC-16 error detection and the HARQ protocol follow the same manner as the MIMO OFDM above.

VI. COMPUTER SIMULATION RESULTS

The throughput performance and the average number of retransmission characteristic for 2×2 MIMO-OFDM NB RCP LDPC coded Type II Hybrid-ARQ with GF(4) and QPSK or with GF(16) and 16QAM are investigated. We compared the performance between with and without relay. The simulation condition is listed in Table II. The simulation results for throughput characteristic are shown in Figure 10 – Figure 13. The simulation results for average number of retransmission are shown in Figure 14 - Figure 17. We have also shown the simulation results for the parallel two relay case of Figure 6 (b) and the serial two relay case of Figure 6 (c) in Figures 18 and 19, respectively.

For QPSK, an LDPC code word on GF(4) is divided into 16 OFDM symbols. As the coding rate of mother LDPC code is 1/2, the former 8 OFDM symbols are the information data symbols and the latter 8 OFDM symbols are the parity check symbols. For the 1st transmission, 8 OFDM symbols made from information data are transmitted from 2 transmit antennas simultaneously using 4 OFDM symbol durations. For the 2nd transmission and thereafter, i.e., retransmission, 4 OFDM symbols made from the parity check symbols are transmitted from 2 transmit antennas simultaneously. In each retransmission, 1 parity check OFDM symbol is transmitted from each antenna using 1 OFDM duration. The coding rate is decreased gradually from 4/5, 4/6, 4/7 to 4/8 for each retransmission. For 16QAM, the coding rate is decreased from 2/2, 2/3 to 2/4. After all the parity check OFDM symbols are transmitted and the coding rate reaches 4/8=1/2, if the errors are still detected at the destination, the whole transmission of the same RCP LDPC code word in the same manner is repeated up to 15 times, which is enough large number to measure the throughput and the average number of retransmission. We call this procedure of decreasing the coding rate from 1 to 1/2 as the one set as shown in Table II. The symbol LLR addition is used at the destination for the repeated reception of the same RCP LDPC code word.

For the comparative scheme, we considered the type I HARQ with the fixed coding rate LDPC code and set the maximum number of retransmissions also to be 15 times.

We compare the throughput performance of type II HARQ with type I HARQ in Figures 10, 11, 12, and 13. As the coding rate of type I HARQ increases, the throughput also increases in the high average E_b / N_0 region. On the other hand, the throughput of type II HARQ approaches to 4 (bps/Hz) and 8 (bps/Hz) in case of QPSK and 16QAM, respectively, in the high E_b / N_0 region. This is because type II HARQ can change the coding rate adaptively and it can use the coding rate of 1 for high Signal to Noise Ratio (SNR) region. The slight decrease of throughput in type II HARQ from 4 (bps/Hz) and 8 (bps/Hz) is due to the use of CRC-16 error detection code. In type II HARQ, however, the parity check packet is sequentially retransmitted in responding to the NACK, so the number of retransmission becomes large compared with the type I HARQ. Also in type II HARQ, the iterative decoding of LDPC code is done for each retransmission of parity check packet, thus the decoding time tends to increase.

TABLE II. SIMULATION CONDITIONS OF NB GF(4) AND GF(16) RCP LDPC CODED TYPE II HYBRID ARQ SCHEME WITH 2×2 MIMO-OFDM

NB LDPC mother code	Size of Galois field	GF(4)	GF(16)
	Size of parity check matrix	(512,1024)	(256,512)
	Average weight	(2.66,5.32)	(2.41,4.82)
	Coding rate	4/8	2/4
Punctured code (efficient code)	Information bit length	1024	
	Coding rate	4/4,4/5,4/6, 4/7,4/8	2/2,2/3,2/4
Max SPA iteration		20	
Transmit and receive antennas		2×2	
Modulation		QPSK	16QAM
Number of OFDM subcarriers		N=64	
CP length (T_s :QAM symbol length)		$T_s / 4$	
Channel model between each transmit and receive antenna		Quasi-static Rayleigh fading with 16 delay paths having equal average power	
Interval of delay paths		$T_s / 64$	
Channel State Information		Known at receiver	
Error detecting code		CRC-16	
Power attenuation exponent		α=3	
Number of retransmission in Type I		15 times	
Number of retransmission in Type II		15 sets	15 sets

Next, we compare the cases with and without relay. When the average receive E_b / N_0 is high, the throughputs with and without relay are almost equal, but when the average receive E_b / N_0 is low, the throughput with relay is higher than without relay. This is because for the high average receive E_b / N_0 region, the destination can receive the packet correctly without retransmission. Accordingly, the relay is not used for this high E_b / N_0 region, so there is no difference between with and without relay. On the other hand, for the region where the average receive E_b / N_0 is low, the transmission from source to destination often fails, but the transmission from relay to destination succeeds with high probability, thus the retransmission is switched from the source to the relay for this low E_b / N_0 region. For the type I HARQ schemes with a relay in Figure 11, Figure 13, Figure 15, and Figure 17, we know that the throughput with a relay is largely improved compared with the one without relay for the region where the average number of retransmission is 1. For this region the throughput of type I HARQ is almost one half of the throughput for high E_b / N_0 region. This means that for this region the transmission is switched from the source to relay and the retransmission from the relay to destination is almost successful. This observation proves that the use of

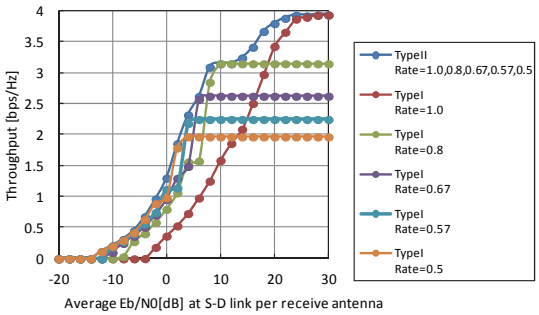


Figure 10. Throughput characteristics of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , QPSK)

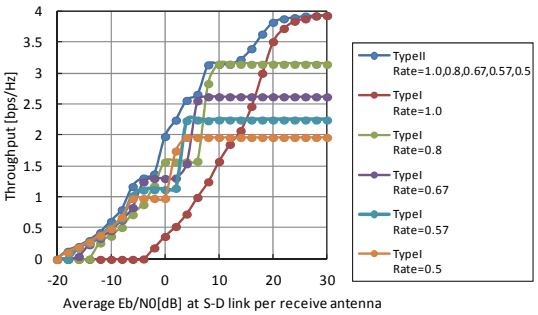


Figure 11. Throughput characteristics of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , QPSK)

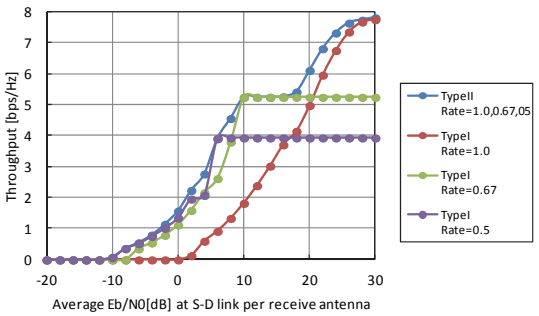


Figure 12. Throughput characteristics of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , 16QAM)

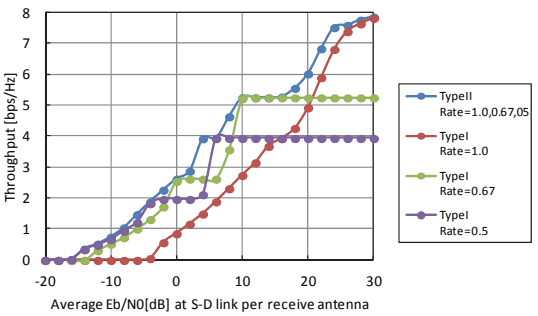


Figure 13. Throughput characteristics of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , 16QAM)

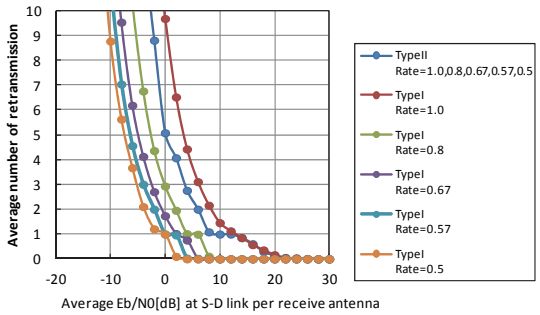


Figure 14. Average number of retransmissions of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , QPSK)

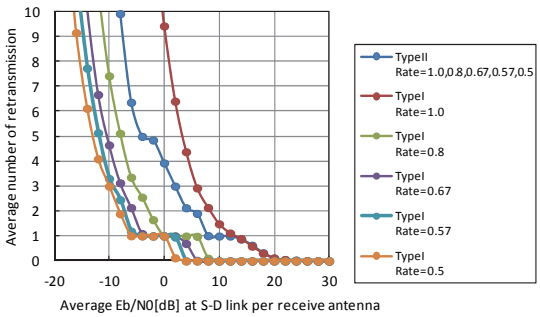


Figure 15. Average number of retransmissions of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , QPSK)

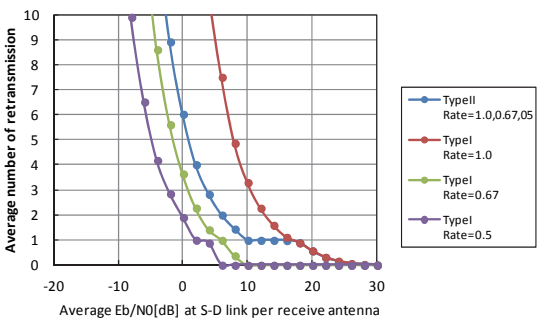


Figure 16. Average number of retransmissions of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , 16QAM)

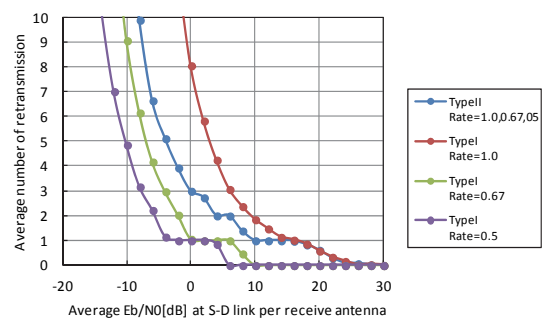


Figure 17. Average number of retransmissions of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , 16QAM)

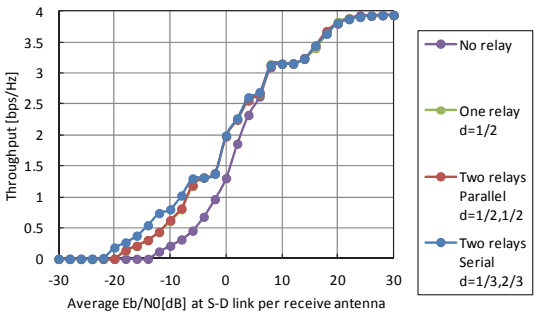


Figure 18. Comparison of throughput characteristics of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy for three relay arrangements (2×2 , QPSK, One relay d=1/2 and two relays parallel d=1/2, 1/2 are overlapped.)

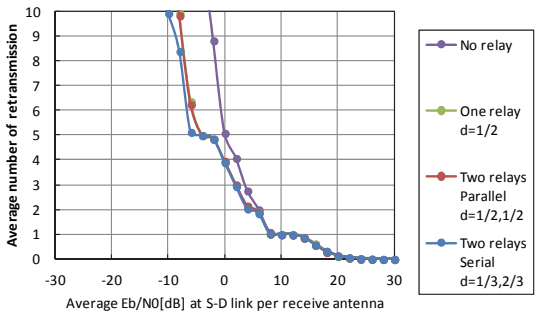


Figure 19. Comparison of average number of retransmissions of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy for three relay arrangements (2×2 , QPSK, One relay d=1/2 and two relays parallel d=1/2, 1/2 are overlapped.)

relay is quite effective in HARQ.

As for the proposed type II HARQ, the throughput is larger than all the type I HARQ schemes and is optimum for all average receive E_b / N_0 values. However, the average number of retransmission becomes larger than the type I ARQ schemes because of the incremental redundancy retransmissions.

As for the relay arrangements in Figure 6, from Figure 18 and Figure 19, we see that the throughput and the average number of retransmission characteristics for the serial arrangement in Figure 6 (c) show the best. The parallel arrangement in Figure 6 (b) exhibits almost the same performance as the one relay case in Figure 6 (a). This observation comes from the fact that the receive power at relay or destination in the serial arrangement becomes larger than the parallel arrangement.

Next, from Figure 20 - Figure 29, we show the throughput and the average number of retransmission characteristics of NB RCP LDPC coded type II HARQ on 2×2 MIMO interleaved SC-FDMA. The simulation conditions are listed in Table III. The simulation results for throughput characteristic are shown in Figure 20, Figure 21, Figure 22, and Figure 23. The simulation results for average number of retransmission are shown Figure 24, Figure 25, Figure 26, and Figure 27. We have shown the simulation results for the parallel two relay case and the serial two relay case in Figures 28 and 29.

When QPSK modulation is employed, one GF(4) LDPC mother code word is divided into 8 packets. As the coding rate of mother LDPC code is 1/2, the former 4 packets contain only information symbols and the latter 4 packets consist of parity check symbols. For the 1st transmission, 4 information packets are transmitted from two antennas. For the 2nd retransmission and after, 1 parity check packets are retransmitted at each retransmission resulting in lowering the coding rate at receiver from 4/5 to 4/6, 4/7, 4/8. After all the parity check packets are retransmitted and the coding rate at destination reaches 1/2, if error is still detected at destination, the same RCP LDPC code transmission is repeated 15 times and each time the symbol LLR's are summed up at destination by symbol LLR addition. We call this procedure of decreasing the coding rate from 1 to 1/2 as the one set. Thus, the total 15 sets of RCP LDPC code word transmission are done before the final discard of RCP LDPC code in case of failure of error correction at destination. As a comparative scheme, we also considered the LDPC coded type I HARQ scheme where the coding rate is fixed for each retransmission. The maximum number of retransmissions is limited to 15 and the symbol LLR addition is employed at the destination.

When 16QAM modulation is employed, one GF(16) LDPC mother code word is divided into 4 packets. The former 2 packets contain only information symbols and the latter 2 packets consist of parity check symbols. The coding rate decreases from 2/3 to 2/4 at each retransmission. After all the packets are retransmitted and the coding rate at destination reaches 1/2, if error is still detected at destination, the same RCP LDPC transmission is repeated 15 times in total, which is the same as the case of QPSK

TABLE III. SIMULATION CONDITIONS OF NB RCP LDPC CODED TYPE II HARQ WITH DF RELAY ON 2×2 MIMO INTERLEAVED SC-FDMA

NB LDPC mother code	Size of Galois field	GF(4)	GF(16)
	Size of parity check matrix	(512,1024)	(256,512)
	Average weight	(2.66,5.32)	(2.41,4.82)
	Coding rate	4/8	2/4
Punctured code (efficient code)	Information bit length	1024	
	Coding rate	4/4,4/5,4/6, 4/7,4/8	2/2,2/3,2/4
Max SPA iteration		20	
Number of users U		4	
Transmit and receive antennas		2×2	
Modulation		QPSK	16QAM
Number of subcarriers / user		N=64	
Number of total subcarriers		M=256	
CP length (T_s :QAM symbol length)		$16 \times (T_s / 4) = 4T_s$	
Channel model between each transmit and receive antenna		Quasi-static Rayleigh fading with 16 delay paths having equal average power	
Interval of delay paths		$T_s / 4$	
Channel State Information		Known at receiver	
Error detecting code		CRC-16	
Power attenuation exponent		α=3	
Number of retransmission in Type I		15 times	
Number of retransmission in Type II		15 sets	15 sets

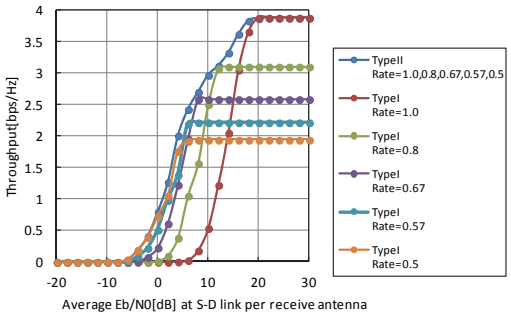


Figure 20. Throughput characteristics of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , QPSK)

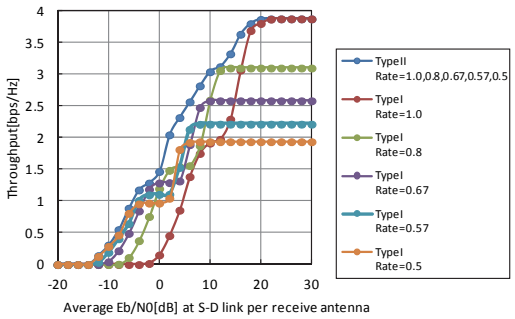


Figure 21. Throughput characteristics of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , QPSK)

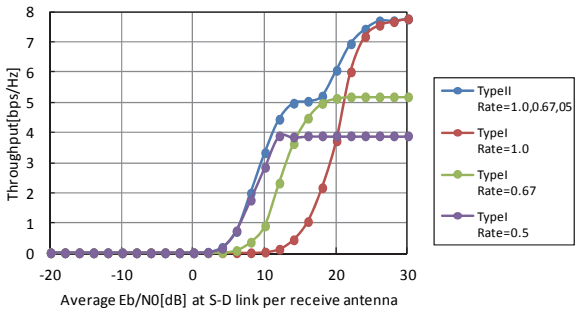


Figure 22. Throughput characteristics of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , 16QAM)

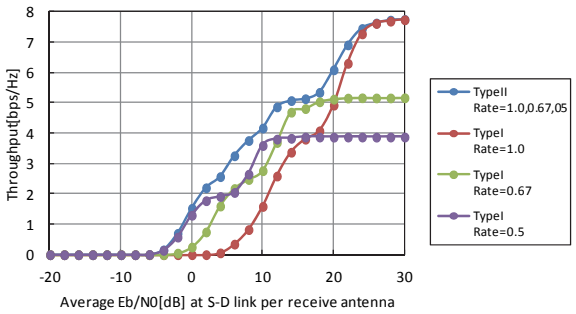


Figure 23. Throughput characteristics of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , 16QAM)

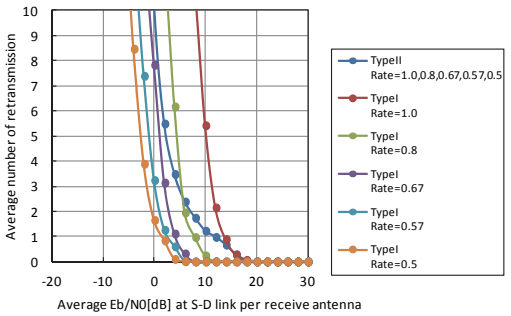


Figure 24. Average number of retransmissions of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , QPSK)

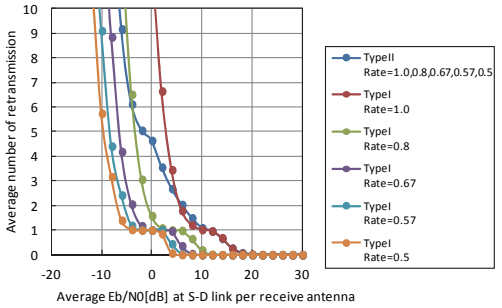


Figure 25. Average number of retransmissions of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , QPSK)

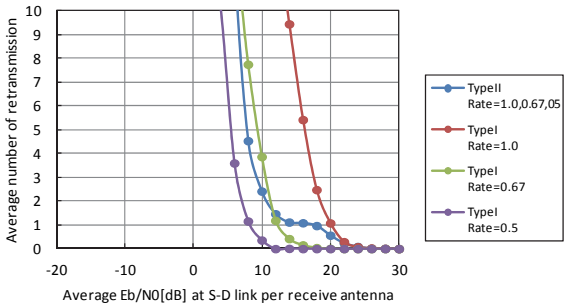


Figure 26. Average number of retransmissions of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (without relay, 2×2 , 16QAM)

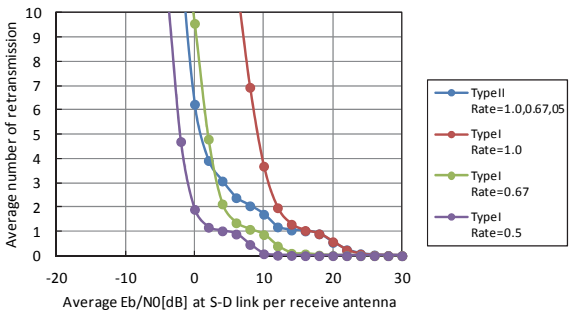


Figure 27. Average number of retransmissions of NB GF(16) RCP LDPC coded type II HARQ scheme with incremental redundancy and type I HARQ with fixed coding rate (with a relay, 2×2 , 16QAM)

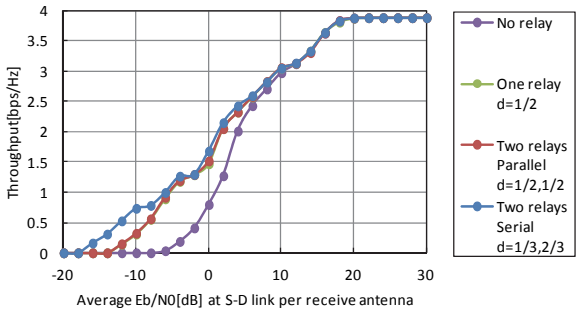


Figure 28. Comparison of throughput characteristics of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy for three relay arrangements (2×2 , QPSK, One relay d=1/2 and two relays parallel d=1/2, 1/2 are overlapped.)

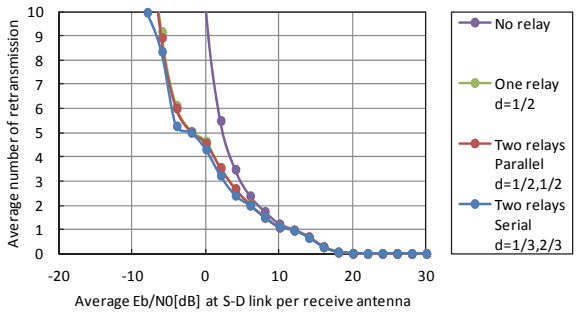


Figure 29. Comparison of average number of retransmissions of NB GF(4) RCP LDPC coded type II HARQ scheme with incremental redundancy for three relay arrangements (2×2 , QPSK, QPSK, One relay d=1/2 and two relays parallel d=1/2, 1/2 are overlapped.)

modulation as mentioned in the above. As for the number of users U , we employed $U = 4$, but any number of users U is available depending on how much subcarriers assigned to each user, i.e., $N = M / U$. Less number of subcarriers will lead to less frequency diversity effect.

For the two parallel relays or two serial relays, we also show the throughput and the average number of retransmission characteristics in Figure 28, and Figure 29, respectively.

Regarding the simulation results, we first compare the type I HARQ with the type II HARQ in case of no relay. The throughput characteristic of type I HARQ saturates in the high average E_b / N_0 region, because the coding rate is fixed. As the coding rate of type I HARQ increases, the throughput also increases in the high average E_b / N_0 region. On the other hand, the throughput of type II HARQ approaches to 4 (bps/Hz) and 8 (bps/Hz) in case of QPSK and 16QAM, respectively, in the high E_b / N_0 region. This is because type II HARQ can change the coding rate adaptively and it can use the coding rate of 1 for high SNR region. The slight decrease of throughput in type II HARQ is due to the use of CRC-16 error detection code. We also observe that for entire E_b / N_0 region, the throughput of type II HARQ is optimized and is superior to type I HARQ. However, the average number of retransmission of type II HARQ is worse than type I HARQ. This is because parity check packets are sent sequentially with several time slots in type II HARQ, while the parity check packet is sent at a

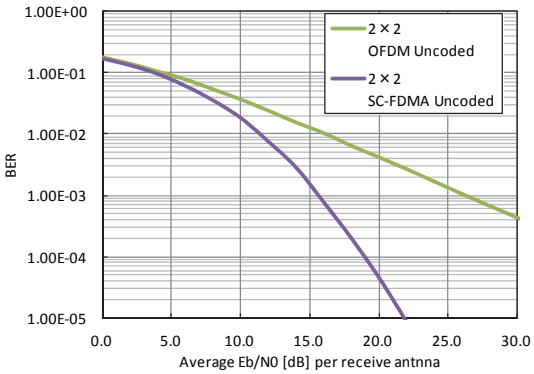


Figure 30. Comparison of BER characteristics between uncoded MIMO OFDM and MIMO interleaved SC-FDMA on 2×2 quasi-static multipath channel

time in type I HARQ.

Next, we compare the case with a relay and without relay. When the average E_b / N_0 is high, the throughput with a relay is the same as without relay. This is because, both in the case with a relay and without relay, the average number of retransmission is almost 0 for the high average E_b / N_0 region, and there makes no difference between the two. On the other hand, when the average E_b / N_0 is low, we see that the throughput and the average number of retransmission characteristics with a relay are much better than the ones without relay. This is because for low average E_b / N_0 region, although the destination frequently fails to decode the code word correctly, the relay succeeds in decoding with high probability. Accordingly, as the retransmission is executed from the relay to the destination instead of the source to the destination, the probability of successful decoding at destination is increased. We also observe that when the number of average retransmission is greater than 0, i.e., the retransmission is done and the total number of transmission is more than 2, the throughput with a relay is largely improved compared with the one without relay. This improvement is observed below $E_b / N_0 \approx 20$ (dB) for GF(4) and QPSK in Figures 20, 21, 24, and 25, and below $E_b / N_0 \approx 25$ (dB) for GF(16) and 16QAM in Figures 22, 23, 26, and 27.

On the relay arrangements in Figure 6, from Figure 28, and Figure 29, we observe that the throughput and the average number of retransmission characteristics for the serial arrangement in Figure 6 (c) show the best performance. The parallel arrangement in Figure 6 (b) exhibits almost the same performance as the one relay case in Figure 6 (a). This observation can be understood from the fact that the receive power at relay or destination in the serial arrangement becomes larger than the parallel arrangement as previously stated in MIMO OFDM.

Finally, we make the comparison between MIMO-OFDM and MIMO SC-FDMA with NB RCP LDPC coded type II HARQ. Almost the same observations we see between the two modulations. On the performance difference between the two modulations, we observe in Figure 10, Figure 14, Figure 20, and Figure 24 that the throughput performance and the average number of

retransmission of MIMO interleaved SC-FDMA are better than the MIMO OFDM above $E_b / N_0 = 20$ (dB). This seems to be caused from the difference of BER performance between the two modulations. We show the BER characteristics of uncoded MIMO OFDM and uncoded MIMO interleaved SC-FDMA in Figure 30 simulated under the same simulation conditions of Table II and Table III. From Figure 30, we know the BER of MIMO interleaved SC-FDMA is better than the MIMO OFDM above $E_b / N_0 = 20$ (dB). Accordingly, the throughput and the average number of retransmission of MIMO interleaved SC-FDMA are superior to the MIMO OFDM.

In all the simulations in the above figures, a quasi-static Rayleigh fading channel with 16 delay paths having equal average power is considered and it is basically a static channel and time-invariant. If the mobile time-variant channel is considered and employed for the proposed schemes, further degradation of throughput and average number of retransmission is expected due to the channel estimation error caused by rapid time-varying channel. For future studies, the throughput and the average number of retransmission should be measured in real mobile environment using real test beds.

VII. CONCLUSIONS

In this paper, we have investigated the throughput and the average number of retransmission characteristics of the proposed NB RCP LDPC coded type II HARQ with DF relays using MIMO-OFDM and MIMO interleaved SC-FDMA. We have verified the effectiveness of the proposed scheme through computer simulation. In the proposed scheme, for the first transmission, only uncoded information packet is transmitted to both for relay and destination. If error is detected at destination, parity check packets are retransmitted for the 2nd and the subsequent retransmissions. The error correction decoding is done both at relay and destination. When the destination fails in decoding, but the relay succeeds, the relay replaces the source hereafter. The relay retransmits the remaining parity check packets with incremental redundancy instead of source. The destination receives the parity check packets till the coding rate reaches 1/2. We made clear that by using DF relays the throughput and the average number of retransmission characteristics are improved especially for low receive SNR region.

ACKNOWLEDGEMENT

This study is partially supported by the Grants-in-Aid for Scientific Research 15K06059 of Japan and the Sharp Corporation.

REFERENCES

[1] T. Hamada and Y. Iwanami, “On Type II Hybrid-ARQ with Decode and Forward Relay using Non-Binary Rate-Compatible Punctured LDPC Code on MIMO SC-FDMA up-link,” ICWMC 2014, pp. 112-117, June 2014.

[2] H. Tanaka and Y. Iwanami, “On Throughput Characteristics of Type II Hybrid-ARQ with Decode and Forward Relay using Non-Binary Rate-Compatible Punctured LDPC Codes,” ICWMC 2012, pp. 272-277, June 2012.

[3] D. Declercq and M. Fossorier, “Decoding algorithm for nonbinary LDPC codes over GF(q),” IEEE Transactions on Communications, vol. 55, pp. 633-643, April 2007.

[4] D. Kimura, F. Guilloud, and R. Pyndiah, “Application of non-binary LDPC codes for small packet transmission in vehicle communications,” The 5th International Conference on ITS Telecommunications, pp. 109-112, Brest France, June 2005.

[5] J. Ha, J. Kim, D. Klinc, and S. W. McLaughlin, “Rate-compatible punctured low-density parity-check codes with short block lengths,” IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 728-738, Feb. 2006.

[6] M. Shimotsu, Y. Iwanami, and E. Okamoto, “An LDPC coded adaptive hybrid ARQ scheme with packet combining on MIMO eigen-mode channels,” IEICE Technical Report, RCS2005-37, pp.59-64, June 2005.

[7] Y. Tsuruta, Y. Iwanami, and E. Okamoto, “A Study on LDPC Coded Hybrid-ARQ Using Spatially Multiplexed MIMO-OFDM,” S36-1, 6 pages, WPMC2009, CD-ROM 5 pages, Sept. 2009.

[8] T. Kozawa, Y. Iwanami, E. Okamoto, R. Yamada, and N. Okamoto, “An evaluation on throughputs for Hybrid-ARQ using Non-Binary Rate-Compatible LDPC codes,” The 32nd Symposium on Information Theory and its Applications (SITA2009), F21-3, pp.771-775, Dec. 2009.

[9] T. Kozawa, Y. Iwanami, and E. Okamoto, R. Yamada, and N. Okamoto, “An evaluation on throughput performance for Type II Hybrid-ARQ using non-binary Rate-Compatible-Punctured LDPC codes,” IEICE Transactions on fundamentals, vol. E93, no.11, pp. 2089-2091, November 2010.

[10] D. Gang, R. Kimura, and F. Adachi, “Performance evaluation of RCPT Hybrid ARQ schemes for DS-CDMA mobile radio over frequency selective Rayleigh fading channel,” IEICE Technical Report, RCS2001-280, pp.241-248, March 2002.

[11] D. Declercq, V. Savin, and S. Pfletschinger, “Multi-Relay Cooperative NB-LDPC Coding with Non-Binary Repetition Codes,” ICN 2012, pp. 205-214, March 2012.

[12] J. Laneman, D. Tse, and G. Wornell, “Cooperative diversity in wireless networks: efficient protocols and outage behaviour,” IEEE Transactions on Information Theory, vol. 50, no. 12, pp. 3062-3080, Dec. 2004.

[13] A. B. A. Aziz and Y. Iwanami, “A simple symbol estimation for soft information relaying in cooperative relay channels,” Int. Journal of Commun. Networks and Systems (IJCNS), Scientific Research Publishing, Vol. 4, No. 9, pp.568-577, Sept. 2011.

[14] H. G. Myung, J. Lim, and D. J. Goodman, “Single Carrier FDMA for Uplink Wireless Transmission,” IEEE Vehicular Technology Magazine, pp. 30-38, Sept. 2006.

Interference Avoidance Routing Strategy in Cognitive Radio Networks

Minh Thao Quach, Francine Krief
LaBRI, University of Bordeaux
Talence, France
Email: quach@labri.fr and krief@labri.fr

Mohamed Aymen Chalouf
IRISA, University of Rennes
Rennes, France
Email: mohamed-aymen.chalouf@irisa.fr

***Abstract*—Co-existence between legacy wireless infrastructure and a cognitive radio network has been attracting the research community. However, many challenges arise due to several difficulties, such as how to leverage current deployment for a secondary network but guarantee no interference to the primary network. This work illustrates a specific coexistence deployment in which a primary coverage reception overlaps with a cognitive transmitter's reception zone. We show that typically, a large overlap zone causes high interference; however, the interference level is lower when the node density is minor. Fuzzy logic is used to combine observed factors of the wireless environment (e.g., area overlapping and primary receiver density) to estimate interference level to primary receivers. The computed results reflect the precise impact that may occur when a cognitive radio communication is operating nearby. Interference level is retrieved by the routing engine and becomes a routing metric alternative to the hopcount metric for our routing proposal, which leverages Dynamic MANET On Demand for cognitive radio networks. In this paper, we detail the proposed routing idea, which promotes a cross-layered routing design for cognitive radio networks and incorporates observed environment information to prevent severe interference with primary networks.**

Keywords–Cognitive radio; overlap region; prediction model; fuzzy logic; interference avoidance.

I. INTRODUCTION

The objective of this paper is to define a routing strategy which avoids interference in cognitive radio networks (CRNs). This strategy is essentially based on our previous work [1]. To make the readers easy to follow, we present the reason why CRNs were introduced. Due to the limitation of fixed broadband and frequency range allocation, there is a huge need for enhancing radio resource usage using cognitive radio technologies, especially for the rural areas. The federal communications commission (FCC) has been conducting data collection and evaluating the communication needs of rural communities since 2008 in the US [2]–[4]. Expanding broadband deployment for rural areas is critical, however their full impact has not been yet realized and it would be interesting to measure this impact.

As illustrated in Figure 1 [5], CRN infrastructure can coexist with an existing primary system including the primary transmitter base station and its primary receivers. Again, evaluating the impact with this co-existing scheme should be studied since it has not been realized thoroughly. To do so, we are interested in studying the reception overlap between the secondary transmitter and primary transmitter and observing its impact on the primary receivers.

Eliminating interruptions during an operation is one of the key properties of reliable applications or services. In CRNs,

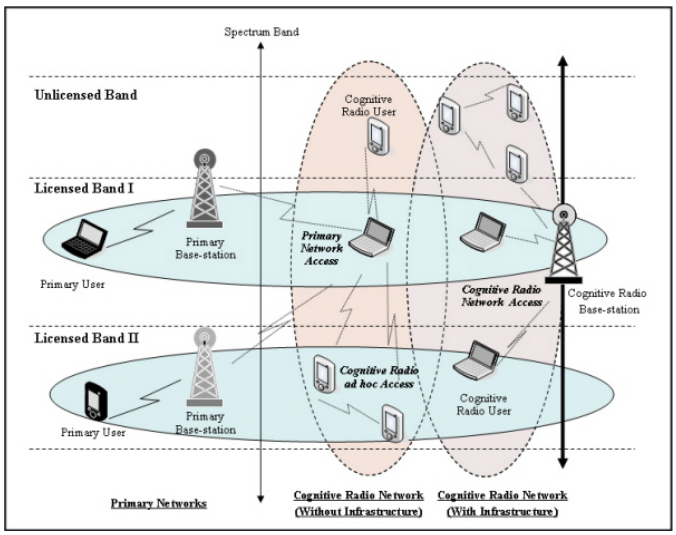


Figure 1. Cognitive radio networks reference architecture

interruptions obviously happen when the transmission of a primary radio (PR) forces cognitive radios (CRs) to vacate on the selected band. On the other hand, when an overlap exists in the coverage area between the CR and PR transmitters, the undesired interference generated by the secondary transmitters on PR communications becomes difficult to control. For this reason, it is necessary to accurately characterize this overlap and its effect on every available channel for better spectrum selection.

We define an overlap region as an area where a PR emitter's signal meets a CR emitter's signal. In case of coexistence, this area is vulnerable to the operating primary receivers within the region. When it comes to protecting PR receivers' communication, avoiding this vulnerable area is one of the options. However, since these receivers are not always operating continuously on this area, the spare spectrum within this area can be exploited and used by CR nodes. Considering the line-of-sight propagation model in rural areas, we define an overlap in CRN context as being a geometrical overlapping area between two circles that were formed by the signal of the PR and CR transmitters. As illustrated in Figure 2, the PR transmitter overlaps with two other CRs that creates two overlap regions, overlap region 1 and overlap region 2 respectively.

We rely on the position of the transmitters and their transmission ranges to identify the encountered overlap situations. Different overlapping cases have different effects on

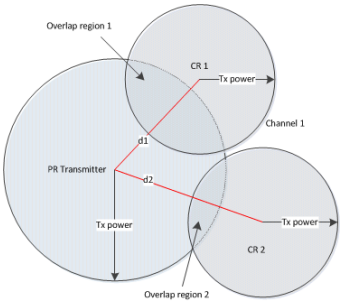


Figure 2. Overlap in cognitive radio networks definition

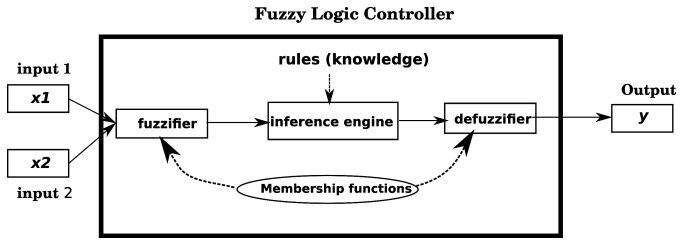


Figure 3. General fuzzy logic system

the primary receivers. Therefore, characterizing all these cases provides better vision on designing an interference-free CR communication scheme. For instance in a rural area, a CRN overlaps with TV transmission signals; the overlap area size could be variable. Assuming that a smaller overlap region leads to less impact on possible TV receivers, this CRN can operate on the area that has small overlapping area. As long as the impact on the receivers is observed, the interference can be avoided.

In a CRN, a CR node makes decisions based on its own observed information even though this knowledge may be incomplete. Fuzzy logic, however, can yield useful outputs with incomplete, approximate or vague information (e.g., low or high interference, sufficient or insufficient available radio resources). Furthermore, fuzzy logic does not require too complicated computation since the calculation is mostly based on if-then-else rules. Hence, we can use fuzzy logic in real-time cognitive radio applications for which the response time is crucial to the system performance [6]. Due to its simplicity, flexibility, and if-then-else rules composition, processing time for fuzzy logic is minor.

Fuzzy logic introduces a logic theory that was developed to generalise 'true' and 'false' values to any value between 0 and 1 [7]. It also presents the approximate knowledge, which may be difficult to express by conventional crisp methods (i.e., bivalent set theory). A fuzzy logic system with two inputs and one output is described in Figure 3. The fuzzy sets are sets of unsharp boundaries in which membership is a matter of degree (in range of 0 to 1). For instance, a fuzzy set of $weekend$ may contain half of Friday, Saturday and Sunday and a set of $weekdays$ may contain from Monday to the first half of Friday. So, Friday can exist in both sets with distinctive degrees. To identify the degree of these variables, a membership function is used to imply the related information. The membership function assigns a value in the interval $[0, 1]$ to a fuzzy variable, denoted by $\mu(weekend(day))$, where $weekend$ is a fuzzy set, and day is a fuzzy variable.

Input crisp values are fuzzified to produce appropriate linguistic values according to defined membership functions. Then, the inference engine will extract the associated outputs based on the defined rules. These outputs are fuzzified based on output membership functions. Finally, fuzzified outputs are aggregated into a single crisp value by the defuzzifier.

The output can be used to investigate how the routing layer reacts and makes the right decisions to maximise spectrum resources while avoiding interference with the primary receivers. For instance, a CR can operate within an area which has high overlap size but low number of operating primary receivers. We apply fuzzy logic to determine the overlap size and the probability of operating primary receivers (e.g., low or high).

In the literature, we did not find any explicit solution or proposal that is sufficient to protect the PR when coexistence happens. There is always a trade-off between the coexistence and interference. In our previous works [8], [9], we graphically observed and estimated how the PRs get impacted when overlap happens along with a proposal of using the stochastic model Grey and Kalman filter to predict the potential PRs that may be impacted. From these observations, we proposed a method to combine these two factors using fuzzy logic, which was detailed in [1]. We explain this again in this article so that the readers understand the proposed routing framework.

The rest of the paper is structured as follows. A literature review of routing solutions in CRN is provided in Section II. Then we briefly present the overlap calculation and the node density estimation method in Section III. The method to compose these two factors into an interference level, in which we estimate this level using fuzzy logic and a routing proposal, is in Section IV and V. Overall, we conclude our work in Section VI.

II. Related Work

Network layer plays an important role in computer network communication, essentially, routing protocol to establish a path between a sender and a receiver. Routing in a wireless network is always an interesting topic in the research community. In CRNs, the routing problem imposes a great challenge due to the dynamic spectrum access nature. If two neighboring nodes do not have a common accessible channel, or they have a common channel but do not tune to the same frequency, the communication in this case is infeasible and no route can ever be established. In CRNs, topology construction includes spectrum detection, neighbor discovery, and topology management. In some circumstances, a routing decision also depends on the required Quality of Service (QoS) from upper layer and also from the control information from the lower layers (PHY and MAC).

Due to the unique function of a CRN, intermittent link is the first challenge that any routing solution needs to tackle. To be specific, routing challenges in CRNs include spectrum awareness, quality route discovery process, and route maintenance/reparation mechanism [10]. Researchers hence have to keep in mind the fact that they will not have any pre-allocation spectrum access for the routing module, so the routing algorithm has to accommodate the change of the environment.

However, the algorithm also has to satisfy the basic network performance [11]. Nevertheless, an inter-dependence solution between route selection and spectrum management could be an appropriate approach to resolve the spectrum awareness issue [12].

Routing over the open spectrum environment is a fundamental issue, especially in dealing with multi-hop CRNs. Several routing solutions were proposed but no general routing solution exists [13]. Again, the challenge is how to ensure radio resources for cognitive transmission while guaranteeing the service for all on-going PR communications over the exploited channels on the whole path. Regarding service-wise perspective, the question is how many possible services can be provided to the end users in the secondary networks. Overall, any routing solution in CRNs should always be aware of the potentially available spectrum that may be provided by the sensing function locally or globally. Research by Cesana, Cuomo, and Ekici [10] classified two main classes of routing algorithm, full spectrum knowledge and local spectrum knowledge, which consider global and local spectrum information among CRs as the main criteria for routing algorithm classification.

Cheng et al. [14] proposed a spectrum-aware routing solution that selects a route according to the switching delay among channels and backoff delay within a channel based on the spectrum information provided. Also, research by Liu, Cai, and Shen [15], another spectrum-aware routing solution, suggested coupling spectrum sensing and spectrum sharing in multichannel for multi-hop routing. Based on the location information and channel statistic, a CR selects the relay hop and adapts its own transmission to dynamic spectrum access opportunities in its neighbor. The authors proposed a routing metric that encountered the throughput called cognitive transport throughput as the main metric for the protocol. This metric was used to capture the dynamic change from sensing information and evaluate potential gain of each relay hop. In term of spectrum-aware solutions, Zhu, Akyildiz, and Kuo [16] also built a spectrum tree based as on-demand routing solution for Cognitive Radio Ad-Hoc Networks (CRAHNs). The global sensing and sharing information were utilized to build a tree with distinguished levels of available spectrum on each band. They suggested cooperating this information from the spectrum decision and routing selection process to produce a metric and adapt an on-demand routing protocol for CRAHNs. Other on-demand routing adaptations for CRNs were also introduced in [17], [18].

Some other works that also take into account the cooperation between spectrum sensing and the routing module were introduced in [19], [20]. Xin, Xie, and Shen [19] proposed a graph-based solution that associated spectrum sensing decisions with the radio interfaces of each node in the network to assign specific spectrum opportunities to the radio interfaces, while Krishnamurthy et al. [20] modified MAC layer configurations to determine a common set of channels to facilitate communication among the nodes. The topology so is formed according to this common set of discovered channels. This solution indeed used the global spectrum knowledge to accommodate the routing algorithm. Additionally, the physical location of each node was also disclosed among the node to provide the global view of the network topology.

Once the knowledge of the surroundings is partially learned, Guan et al. [21] proposed a prediction-based middleware between the network layer and lower layers. The authors studied topology control and routing issues in Cognitive Radio MANETs (CR-MANETs) and built a middleware-like cross-layer module to provision cognition capacity to do routing for CR-MANETs. The work aimed to capture the dynamic change of topology and potentially construct an efficient and reliable topology. Indeed, the solution is the inter-dependence component as mentioned between MAC layer and routing layer, and this component incorporates sensing statistics to predict and provide information on the available duration of link to routing components. Other solutions that deal with opportunistic networks include an opportunistic access routing solution in [13], where routing metric was rendered based on demanded QoS in cooperation with the channels' access opportunities.

Location awareness is another aspect of concern in CRN routing. A geographic forwarding based Spectrum Aware Routing protocol for Cognitive ad-Hoc networks (SEARCH) is also a location awareness routing solution in [22] that joined undertaken paths to completely avoid the PR's region, while routing to protect the PR devices within this area. Similarly, Habak et al. [23] suggested location-aware routing solutions could provide better protection for PR devices from CRNs' communication. Also, a routing solution with consideration about overlapping was proposed in [24]. PR receivers protection was studied in this work as the main purpose. The routing mechanism ensured a perfect protection for PRs by selecting routes that avoid any overlap between PR and secondary radio coverage. However, the resources of the overlap region may be usable when the PR receivers are inactive or non-existent.

Even though the research community has been spending considerable resources and effort to resolve routing challenges for CRNs, there are still no routing standards that could overcome all the challenges. Routing issues encountered in CRN design can be varied. Each hop has its own different expectation of available resources (channels, frequency range, power level, interference level, etc.) and allocation time. Therefore, most of the existing solutions treat routing designs as a cross-layered problem with cooperation between a lower layer, e.g., to acquire sensing information and to render potential resources, and the routing mechanism on the network layer to establish a path among nodes.

Another perspective that attracts the research community is interference analysis in CRNs. The investigation of interference analysis takes a critical role to provide input to design various network parameters to guarantee certain performance for the primary users [25]. Many studies have been carried out in [26]–[30]. Hossain et al. [25] have studied and classified two types of interference configuration, network with beacon and network with primary exclusive regions.

Assuming that in a circular network the CRs are uniformly distributed with a constant density λ, the interference generated by these CRs depends on their locations and on the random channel fading, so this type of configuration causes random interference [31]. In the network with beacons where CRs capture and detect beacons sent from PRs, the CRs try to avoid transmitting in the next duration after successfully detecting these beacons. The PRs' communication is hence safe from

being perturbated. However, again the channel fading problem could make these CRs misdetect the beacons. Therefore, a beacon detection threshold becomes on a crucial parameter to design a CRN that limits the impact on the PR's operation. In the network with primary exclusive regions, the PR transmitter's exclusive region is better avoided by any CRs within the area since PR receivers within this area are passive devices and they might be affected if any of the nearby CR operates.

Also depending on the locations of CRs and PRs, our work studies the network environment's observations from the CRs point of view to evaluate the impact, which an on-going CRN may cause to the primary network. Our observations take into account the inclusive zone between PRs' and CRs' networks. This awareness is then incorporated into a metric called interference level that reflects the importance of the impact. Our proposed routing solution uses the interference level as the crucial metric on select the route that minimizes the impact to the primary networks that is unnecessarily completely avoided when inactive or non-existing PR receivers are detected. Our routing proposition is also a cross-layered approach where the network layer accesses and retrieves physical information directly from the physical layer via a management controller (see Section V-B). In the following sections, we detail the different parts of our work: overlap observation, impact estimation and routing proposal.

III. Overlap Region and Node Density Observations

In this section, we present the overlap region observation and node density estimation approach, which are the essential piece of information to envisage possible interference on the PR's system when CR's system is operating.

A. Overlap Region Observation

We argue that when the reception area of a CR emitter and the reception area of a PR emitter overlap, it produces unavoidable effects on the primary system, especially to the PR receivers. This observation was mentioned in [8]. However, we also proved that not only the overlap size but also the number of existing primary receivers cause the impact [9]. Our approach takes into account the overlap ratio and node density probability as two main factors. The ratio of the overlap size to the overall size of the PR emitter's disk is named overlap ratio, while the node density probability is the probability of possible node density within this PR emitter's disk.

The realistic overlap case highlighted in Figure 4 and Figure 5 is obtained when the following condition is satisfied. We study the possible impact of these general cases.

$$d < R_P + R_C \quad and \quad d > R_P \quad and \quad d > R_C. \tag{1}$$

The overlapping region is clearly displayed and by using classical geometry, we calculate the general overlap region between the PR and CR for this case in equation (2),

$$\begin{aligned} A_{Overlap} = & \frac{\theta_C}{2} R_C^2 - R_C^2 |\cos\beta| \cos\frac{\theta_C}{2} \\ & + \frac{\theta_P}{2} R_P^2 - R_C |\cos\beta| R_P \cos\frac{\theta_P}{2}; \end{aligned} \tag{2}$$

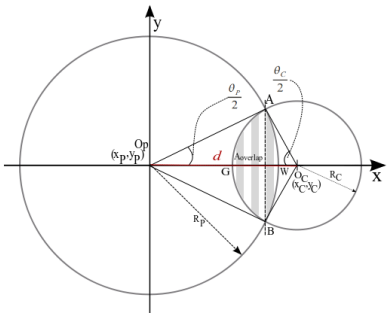


Figure 4. General overlap case

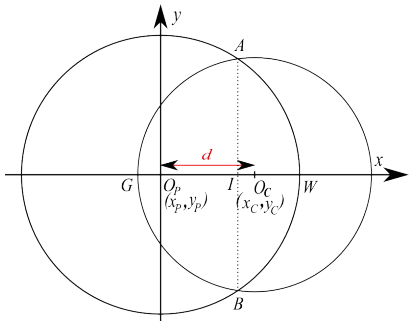


Figure 5. Particular overlapping case between PR Transmitter and a CR Node

where θ_C and θ_P are the angles formed at O_C and O_P with points A and B, respectively, whereas β is an intermediate variable for our computation given,

$$\begin{aligned} \sin\beta &= \frac{R_P^2 - \left(R_C^2 + x_C^2 + y_C^2\right)}{2 * R_C * \sqrt{x_C^2 + y_C^2}}; \\ \sin\frac{\theta_C}{2} &= \frac{R_C}{R_P} * |\cos\beta|; \\ \sin\frac{\theta_P}{2} &= |\cos\beta|. \end{aligned}$$

However, the case where $d < R_C < R_P$ shown in Figure 5 is not included in equation (2). This particular situation is captured in equation (3).

$$\begin{aligned} A_{Overlap} = & \frac{\theta_P}{2} R_P^2 - R_C R_P |\cos\beta| \cos\frac{\theta_P}{2} \\ & + \Pi R_C^2 - \frac{\theta_C}{2} R_C^2 \\ & + R_C^2 |\cos\beta| \cos\frac{\theta_P}{2}. \end{aligned} \tag{3}$$

Following up with the experiments, we present some associated results in this section. Figure 6 shows results for a scenario where over a single channel network we modify the overlap size and study its impact on PR receivers. PR receivers were uniformly deployed around the PR emitter. The x−axis illustrates the size of the overlap obtained from the location of the CR sender and calculated based on equations (2) or (3); the y−axis shows the ratio of impacted primary nodes. Lower coverage area in Figure 7 produces almost the same observed results on the single channel experiments.

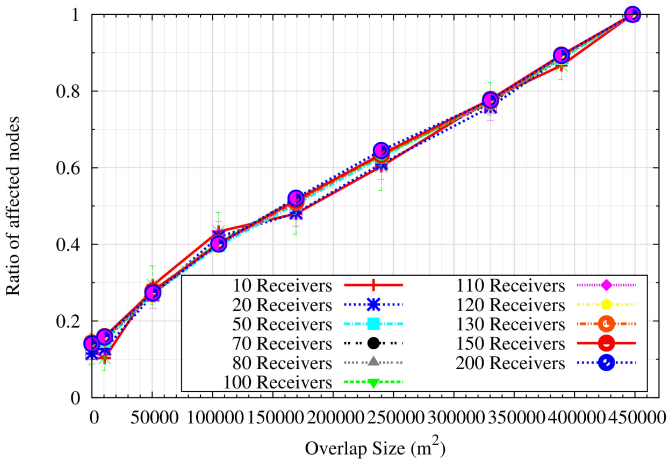


Figure 6. Single Channel - Different Overlap regions - Uniform Distribution

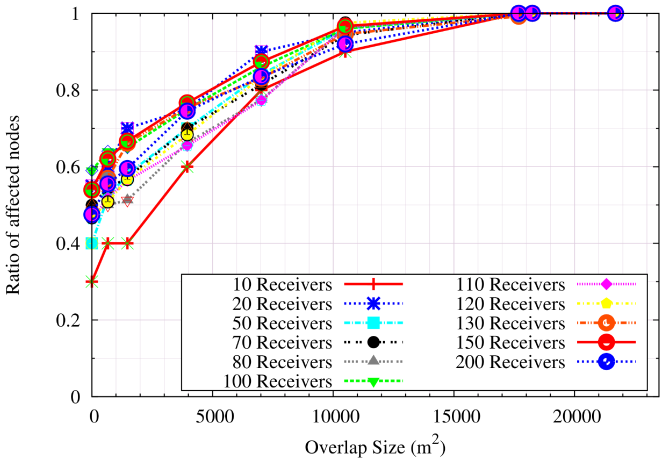


Figure 7. Single Channel - Different Overlap regions, smaller reception zone - Uniform Distribution

Figure 8 corresponds to the second scenario where two channels are available with a PR transmitter observing different overlaps on each channel. In fact, the overlap on channel_1 is about $120000m^2$ while on channel_2 the intersection between transmitter circles is around $500m^2$. In Figure 8, we study the distinctive ratio of affected nodes while modifying the number of PR receivers on both channels.

Again, the experiment is performed with various numbers of PR receivers within the PR transmitter's disks on both channels. The x-axis shows the number of PR receivers while the y-axis illustrates the ratio of nodes being affected by the CR transmitter's signal. This ratio varies from 10% to almost 15% on channel_2 but from almost 34% to almost 40% on channel_1 when the receivers are distributed uniformly within these disks (as in Figure 8).

Though the relationship between overlap sizes and the impact is shown, we performed another experiment on a single channel network with different overlap sizes and arbitrary

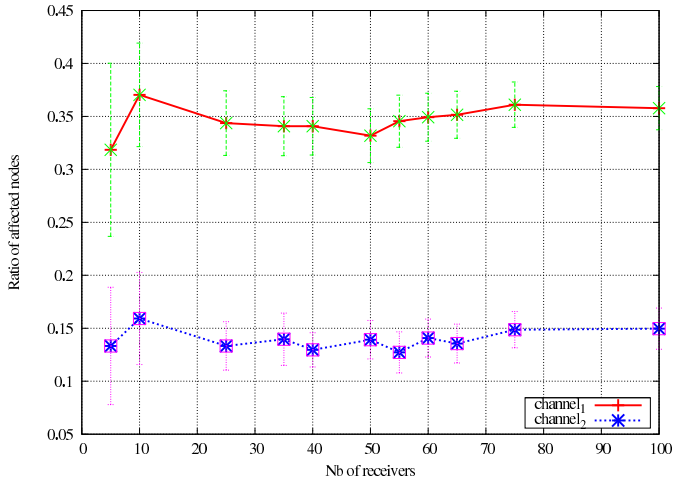


Figure 8. Multi Channel - Different Overlap regions - Uniform Distribution

deployment of PR receivers. In this test, the primary users were deployed further from the vulnerable area. For instance, even when 200 receivers were deployed, the ratio of affected nodes was lower than the deployment of 80 or 50 receivers in Figure 9 and Figure 10, respectively. Hence, envisaging the distribution over the overlap area and applying the proper prediction model to estimate the density of the receivers could be a new approach to protecting the primary receivers in general.

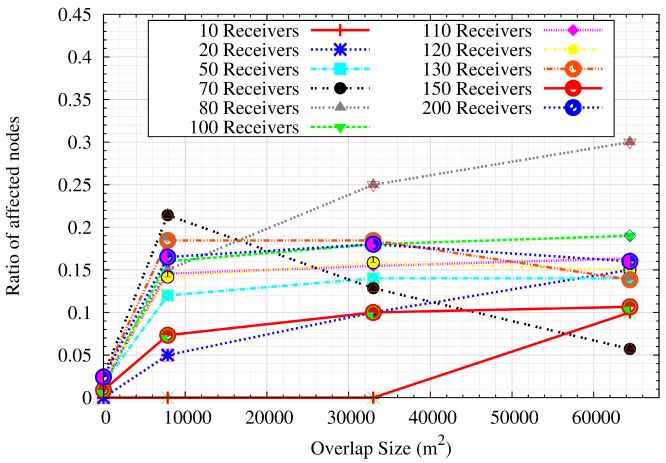


Figure 9. Single Channel - Big Reception Zone - Arbitrary Deployment

In summary, the simulation results reflect the relationship between the overlap area and the impact on the primary receptions with homogeneous distributions of PR receivers. However, positions of PRs may be affected by practical considerations such as obstacles, buildings, mountains, etc., which prevent regions of the coverage area from containing receivers. Therefore, a prediction model to estimate the location of these devices is required.

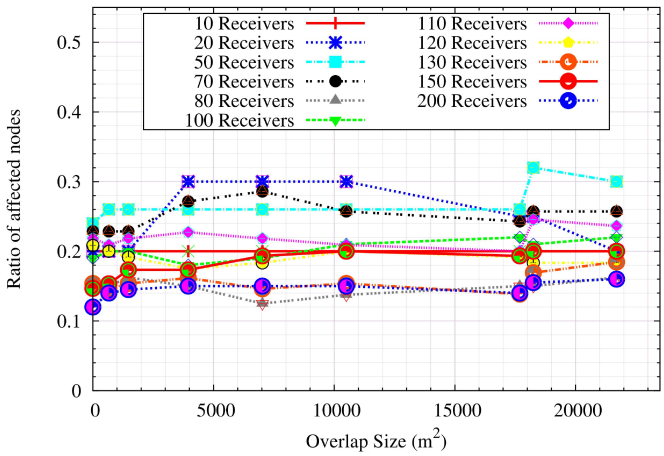


Figure 10. Single Channel - Small Reception Zone - Arbitrary Deployment

B. Node Density Estimation Approach

We now look for a solution for estimating the node density within a specific area. There are two approaches that are taken into account in this work the grey model GM(1, 1) and the Kalman filter. The Grey systems theory [32] is known for the analysis of problems with incomplete or uncertain information. This is an uncertain system in which the information is incomplete and the existing data is partially accurate. Grey system focuses on the uncertainty problems of known samples or inadequate information. Many components of Grey system theory have been built up since the theory was introduced in [33]. These components consist of systems analysis, evaluation, modeling, prediction, decision-making, etc. Our work focuses on the Grey prediction in terms of modeling and predicting on a set of samples.

In the literature, the Grey predictor is robust with respect to noise and lack of modeling information compared to other prediction methods [34]. The Grey model definition is quoted [32]: "Grey models predict the future value of a time series based only on a set of the most recent data depending on the window size of the predictor assuming that all the data used in the models are positive and the sampling frequency of the time series are fixed." One of the most efficient Grey models in real time applications is GM(1, 1). The model GM(1, 1), pronounced as "Grey Model First Order One Variable", is the time series forecasting model in which the model is renewed when new data become available to the prediction model [34].

In next generation networks, e.g., software-defined radio networks or so-called CRNs, the resource exploration process on these devices has very limited information about radio environment. A prediction model that satisfies partial knowledge can be useful. The density of these devices for example is one of the important characteristics that a CR needs to estimate to avoid the maximum perturbation on the primary system. A proposal on the use of the Grey Model in [35] defined a function that facilitates the detection of free-bands by the mobile cognitive radio equipment dedicated to the real-time patient's monitoring. This work basically proposed a predictive strategy, which is based on machine learning techniques combined with the Grey Model system for performing a spectral prediction. Although we characterize the relationship between the overlap area and the effect on the primary system, the primary receivers' density within this vulnerable area is vaguely known. We argue that the Grey Prediction Model is able to estimate the mobile node position.

Firstly, we are interested in predicting the distribution of the PR receivers within a specific area using the forecasting model GM(1, 1) and the Kalman filter. Since we are short of space in this article, we encourage readers to refer to details of the model GM(1, 1) explained in [9], and [35]. The model uses least-squares method to adapt the prediction curve to the original curve. The idea of using the least-squares method is to find the most closest point to the actual value curve/line of the data. In the Grey Model, to ensure the accuracy of the predicted value in the time domain (that reflects the trend of the series) and prediction control variation (the variance of the predicted value and the actual value), the least-squares method is applied to calculate these coefficient factors.

We adapt GM(1, 1) to predict primary receivers on a specific area if a statistical receiver quantity is known. Assuming we have statistical information of the number of active licensed mobile devices in a specific area (note that we focus on the rural area since there are less obstacles in-between). Depending on the time unit of the statistic, we extract the statistic into a time series chain of the number of active licensed devices, which is named "predicted PRs". This is the input series of the Grey Model GM(1, 1). In current experiments, the relationship between transmission power and reception power in free space can be approximated by [36]:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2 L}$$

where P_r and P_t are received and transmitted power, respectively, G_r and G_t denote the antenna gains of receiver and transmitter, L the system loss factors (a.k.a. filter losses and antenna losses), d the distance between the transmitter and the receiver and λ is the wavelength of the transmitted signal in meters. Thus, knowing the PR's positions based on their transmitting/received power by the CR node, the system (i.e., Cognitive Radio emitter) could estimate the changes in their position and the number of PR receivers from the overlap region. According to this information, the CR node could adapt its transmit parameters (power) to minimize the possible impact on PR systems.

Another approach that we consider is the Kalman filter. The Kalman-based estimator provides good results in practice due to its optimality (it minimizes the mean square error of the estimated parameters). The Kalman filter recursively corrects the estimation function associated with the current measured value, current predicted value and current predicted measurement value. For example, suppose we have a prediction function of variable x at time t called, $\hat{x}_t^-$. The current measured value is z_t, the predicted measurement, which is yielded by the predicted value of x at t is ${z^-}_t$, providing that ${z^-}_t = H{x^-}_t$ with H is the related matrix to the measured value at time t.

If the actual value of x is denoted by $\hat{x}_t$, the prediction function is then corrected as follows:

$$\hat{x}_t = \hat{x}_t^- + K(z_t - H\hat{x}_t^-) \quad (4)$$

K is called the Kalman gain associated with the error from the actual measured value and the prediction measurement value, z_t and $z^-{}_t$, respectively. In (4), the composition after K is the correction for the predicted function for estimation of x, denoted by $\hat{x}_t^-$.

We apply the same adaptation and series to perform an experiment with the basic Kalman filter model. The idea behind comparing these two methods is the data smoothing process that the Kalman filter provides. We found that the rendered plots provide similar effects between the original series and the predicted series. The collected results in Figure 11 and Figure 12 show the relationship between the received signal and the distance. The future position is predicted based on this received signal and the current position. These examples (Figure 11 and Figure 12) allow us to estimate the different positions of the PRs in the overlap region and to characterize the overlap region with PRs mobility. The sensitivity range for this case has a limit which approaches $-100dBm$. These results can also be the basis for adaptive transmission power selection.

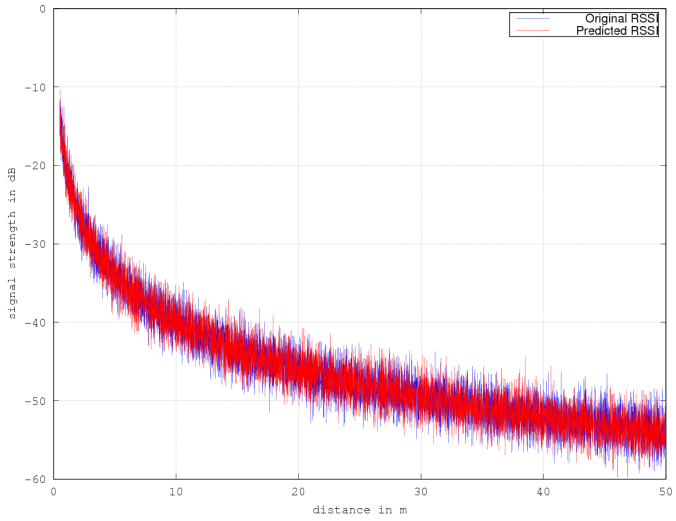


Figure 11. Relation of RSSI and distance with Grey Model GM(1, 1)

IV. Fuzzy Logic-based Interference Level Estimation

From the observations presented in the previous section, we combine overlap ratio and node density probability using fuzzy logic in this section. These parameters are the fuzzy inputs of our fuzzy inference system. A proper implication is applied for each rule listed in the rule table. The result from the implication rule is then aggregated and defuzzified to obtain the final result. This is the degree of impact on the primary system. A CR can consider this degree before using a frequency range when overlap happens.

To interpret the output of antecedents (i.e., the overlap ratio), we use the Mamdani Min Implication rules [7] to extract

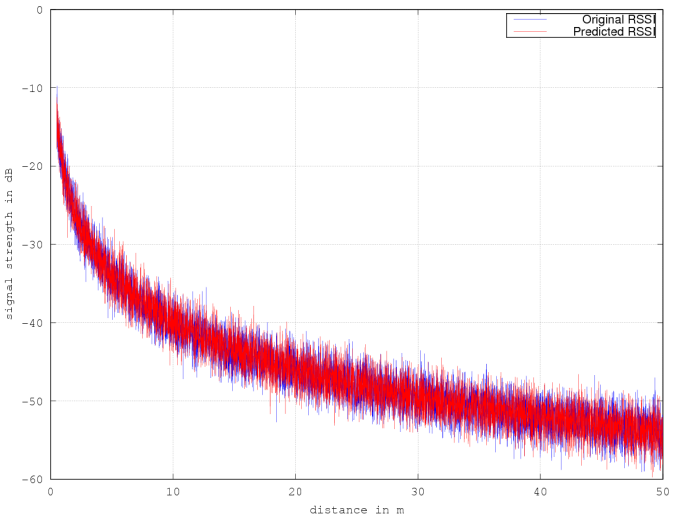


Figure 12. Relation of RSSI and distance with Kalman Filter

the final result for the overlap ratio fuzzy set. For instance, at the intersection of two functions, an *or* operator is used to connect two sets, and the maximum of two membership functions is evaluated for the antecedent part of the fuzzy rules.

$$\begin{aligned} \mu_{OverlapRatio}(x) &= \mu_{Low}(x) \vee \mu_{High}(x) \\ &= max[\mu_{Low}(x), \mu_{High}(x)] \end{aligned} \quad (5)$$

Detailed explanation and implementation to produce the interference level was described in our article [1], hence we only summarize the sample rules table, its inferred statements and the final results of the aggregation process to produce the interference level from the overlap ratio and node density probability.

TABLE I. ENHANCE INTERFERENCE LEVEL RULES TABLE

Index	Overlap Ratio	Density	Interference Level
1	Low	Low	Low
2	Low	Medium	rather Medium
3	High	Medium	somewhat High
4	High	High	High
5	Very High	High	extremely High
6	Very High	Very High	extremely very High

For instance, the above rules infer the following.

- If (Overlap-Ratio is *Low*) or (Density-Ratio is *Low*), then (Interference Level is *Low*) (1)
- If (Overlap-Ratio is *Low*) and (Density-Ratio is *Medium*), then (Interference Level isn't *Low* or rather *medium*) (0.5000)
- If (Overlap-Ratio is *High*) or (Density-Ratio is *Medium*), then (Interference Level is somewhat *High*) (1)
- If (Overlap-Ratio is *High*) or (Density-Ratio is *High*), then (Interference Level is *High*) (1)

- If (Overlap-Ratio is Very *High*) and (Density-Ratio is *High*), then (Interference Level is extremely *High*) (1)
- If (Overlap-Ratio is Very *High*) or (Density-Ratio is Very *High*), then (Interference Level is extremely Very *High*) (1)

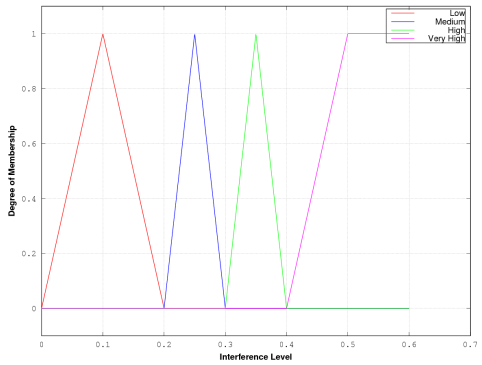


Figure 13. Interference Level membership function

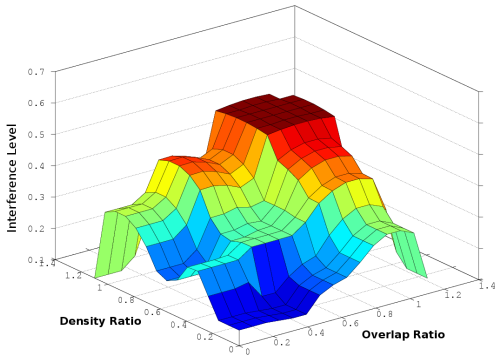


Figure 14. Interference Level Output as function of Overlap and Node Density - Rule set of 16

The fuzzy inference engine combines the rules to obtain the aggregated fuzzy output. The output is the fuzzy set of the interference level that is defined in Figure 13. The Fuzzy controller has to defuzzify this output into crisp values using the centroid method to make the final decisions. Figure 14 shows the system output as a function of 2 variables, overlap ratio and node density, with the rules set of 16 conditional statements.

Table II and Table III show the fuzzified data of the inputs based on their defined membership functions. Table IV shows the output of the inference system. The aggregated values are processed according to the interference membership function in Figure 13. The rules are applied correspondingly in the defuzzification process to produce the final crisp value of the interference level.

More precisely, the Fuzzy Logic Controller (FLC) maps the fuzzified outputs (a.k.a., the output of each linguistic variable of overlap ratio and node density probability) of the inputs to infer the associated consequences. For instance, the inference engine decomposes an input of overlap ratio of 0.25596 into

TABLE II. OVERLAP DEGREE FUZZIFICATION OUTPUT

Index	overlap ratio	$\mu(x)$ Low	$\mu(x)$ Medium	$\mu(x)$High	$\mu(x)$ Very High
1	0.25596	0.29362	0.37305	0	0
2	0.57829	0	0	0.81143	0
3	0.44402	0	0.37322	0.29345	0
4	0.66566	0	0	0.22894	0.26264
5	0.53420	0	0	0.89467	0
6	0.65453	0	0	0.30314	0.21811

TABLE III. DENSITY DEGREE FUZZIFICATION OUTPUT

Index	Density probability	$\mu(x)$ Low	$\mu(x)$ Medium	$\mu(x)$ High	$\mu(x)$ Very High
1	0.08049	1	0	0	0
2	0.19908	0.67281	0	0	0
3	0.66764	0	0	0.21574	0.27056
4	0.81842	0	0	0	0.87370
5	0.28182	0.12118	0.54548	0	0
6	0.57480	0	0	0.83466	0

TABLE IV. INTERFERENCE LEVEL FUZZIFICATION DATA

Index	Overlap input	Density input	Interference Level	Crisp Value
1	0.25596	0.08049	Low	0.14236
2	0.57829	0.19908	Medium	0.21464
3	0.44402	0.66764	High	0.39495
4	0.66566	0.81842	Very High	0.47103
5	0.53163	0.74970	Very High	0.40053
6	0.65453	0.57480	High	0.37930

$\mu(low)$ at 0.29362 and input of density probability of 0.08049 into $\mu(low)$ at 1. This composition matches the first rule in Table I - *If (Overlap-Ratio is Low) or (Density-Ratio is Low), then (Interference Level is Low).*

Figure 14 shows the system output as a function of 2 variables, overlap ratio and node density, and the output of interference level. Each of the variables presented in Figure 14 reflects both the linguistic and also the crisp values. For instance, when the overlap degree and density degree are zeros, the interference is also almost zero (coded with dark-blue area). While the overlap ratio is high (over 60% certainty) and the node density degree is low (below 20%), the interference level is low with low probability (the blue area). This explains the fact that the interference level still strongly connects to the 20% degree of node density that might be affected by the big overlap ratio. The higher the node density, the higher the interference level with high probability.

This interference level can be used in routing in CRNs, which we present in the next section.

V. ROUTING DESIGN PROPOSITION

To get the picture of current work for routing in mobile networks, we go through some existing approaches for routing solutions. There exist three types of routing protocol designs for mobile networks: proactive, reactive and hybrid. Proactive routing protocol are protocols that require node establishes a route to all destinations regardless of the applications' demand. In reactive routing protocol, on the other way round, a node only initiates routing explorations upon applications' demand. Hybrid protocols merge the advantages of both proactive and reactive features.

On Link State Routing (OLSR) is a well-known proactive routing protocol while Ad-hoc On-demand Distance Vector (AODV) and Dynamic Source Routing (DSR) are common

reactive routing protocols. However, which routing protocol can work efficiently in CRNs and how much adaptation could be considered? There are some considered facts in cognitive radio ad-hoc network routing design, as follows. Routing protocol in CRNs should

- avoid periodic messages since redundant messages sometimes cause collision and confusion,
- avoid acknowledgement over the CR link, so that the overhead of the protocol is not too big,
- detect primary users' existence so that the routing exploration process does not cause serious impact to the primary networks.

We first look at some current research that studies the performance of reactive routing protocols. In literature, much work has been conducted to verify the efficiency of these reactive protocols but obtained results are controversial. We present some of the work in the following.

Johnson and David [37] introduced DSR in 1996 and the protocol was then standardized in RFC 4728 [38] in 2007. DSR was designed to quickly adapt to routing changes when a mobile host moves frequently in an ad-hoc network. A route establishment process can only initiate on the desire of the data sender - when a host wants to send a data packet to a destination.

AODV was introduced in 1999 by C. E. Perkins, E. Belding-Royer, and E. M. Royer [39] and was standardized in RFC 3561 [40]. The original design of the protocol has been vastly used by the research community. Many spin-off of AODV have been proposed as the routing protocol for different types of mobile networks.

In principle, AODV is a table-driven routing protocol while DSR is cache-driven. AODV has route maintenance features while DSR lacks route maintaining function. DSR has to re-initiate the routing exploration process when there is a link breakage. Consequently, data communication may be interrupted.

AODV's creator and his co-authors conducted a comparison regarding the performance of these two routing protocols in ad-hoc modes in [41]. Packet delivery fraction (or ratio in some of the recent articles), average end-to-end delay and normalized routing load were the main metrics used to measure the performance of AODV and DSR. The performance results show that DSR outperforms AODV in small network topology and lower mobility. However, AODV handles changes better - i.e., nodes move faster - than DSR does in dynamic networks thanks to the tracking mechanism. Basically, since AODV keeps track of actively used routes, multiple actively used destinations can hence be searched using a single route discovery flood. This mechanism helps to control routing load over the medium, so that the overhead of the protocol seems to be lower than that of DSR. However, the overhead may come from MAC instead causing the routing load increases.

Recently, Jain et al. redo the comparison of these two protocols on the basis of propagation path loss models [42] while Manickam et al. in [43] examine the performance of these two protocols similarly to the work of Perkins et al. in [41].

TABLE V. ROUTING PROTOCOL COMPARISON TABLE

Protocol	Type	Metric	Periodic message	Acknowledgement message	Applicable network
DSR	Reactive	Shortest path	Yes	Yes	Ad-hoc but less dynamic than AODV
OLSR	Proactive	Shortest path	Yes	Yes	MANETs
AODV	Reactive	Hop count	Yes	Yes	MANETs
DYMO	Hybrid	Hop count with alternative metrics	No	Optional	VANETs and MANETs

The concluded results suggest similar performance to that described in [41]. Yet in [42], [43], DSR has lower end-to-end delay than AODV since the exploring path from source to sink is attached in every routing request. By reversing the enclosed path, the sink does not take much time to travel back to the source, so that the processing delay is reduced. Regarding the packet delivery fraction, AODV performs consistently better than DSR does in both [41], [43].

Regarding the performance of a proactive and a reactive routing protocol, we found a study of Sagar et al. [44] in which OLSR is compared with Dynamic MANET On-demand (DYMO) or AODVv2 routing protocol. DYMO was designed to deal with modern day dynamic ad-hoc network topologies. It works as a reactive and also a proactive routing protocol. Initiating discovering routes when it is active or looking up for a route on demand are two different configurations of DYMO. In fact, the protocol is intended for use by mobile routers in wireless, multihop networks. DYMO operates very similarly to AODV, but requires only the most basic route discovery and maintenance procedures [45]. DYMO has also been built with enhancements. Most of the optimizations available in AODV should be applicable to DYMO as well. Sagar et al. [44] have evaluated DYMO and OLSR on MANETs and VANETs. The results suggest that DYMO performs better than OLSR does especially in VANETs where the mobility of mobile nodes are dynamic and the surrounding is quickly changed. Other analyse on the performance of DYMO are also discussed in [46], [47]. All these works concluded that DYMO outperformed the existing routing protocols for MANET in terms of packet delivery, delay and compatible throughput. However, in [45] the authors suggest that DYMO assumes the link between mobile nodes are bidirectional. More precisely, this routing relies on the accuracy of the information provided from the MAC layer.

Overall, a proactive routing protocol requires that a route is established and maintained before any data communication is needed. Therefore, it may not be suitable for CRNs due to the overhead of the control messages all over the channels. Consequently, the resources of a cognitive node exploits may just be used for routing the control message. However, with a reactive routing protocol, the route exploration process is only initiated when a node wants to start a data communication, though route maintenance should be also considered as part of the design. We want to leverage current design with minimal customization. Therefore, adopting both reactive and proactive approaches appears to be reasonable. The DYMO protocol is a potential candidate for this purpose. We will briefly discuss the protocol and its essential operations; e.g., route discovery and route maintenance.

A. DYMO Operation

The Dynamic MANET On-demand DYMO routing protocol has been developed and defined as an IETF specification

(Internet draft version 4.0) [48]. This is a successor of the AODV routing protocol, which was proposed by the same main author. DYMO has AODV's attributions, which can work basically in mobile ad-hoc network mode. For instance, route discovery is only activated when needed. Loop prevention is done by attaching the sequence number to each routing message. Supposedly, a node only creates a route when it has a demand to send data messages to the target node. Therefore, the intermediate nodes maintain and detect topology changes while listening to these exchanging data messages. No HELLO packets are needed to distinguish AODV, so the change is detected faster and the overhead of sending extra control messages is lower. However, some requirements of DYMO should be taken into account to guarantee the best performance of this protocol. The protocol assumes to work under the bidirectional link. It hence relies on the reliability of the lower layer to provide essential environment state information. We will discuss this relation in later Sections. Firstly, we briefly describe the main operations of DYMO in MANET.

Route Discovery: When a mobile node wants to establish a path to communicate to another node in the network, this node has to discover the path to reach its desired destination. To do so, this node, called the source, generates a route request and transmits this request to any of its known neighbors. Each of these neighbors, including the source and the target node, will maintain the active discovered link via a routing table, which is created right after the discovery process finishes. Each entry contains basic information for routing such as the addresses of source and sink, sequence number, hop-count, next hop address, next interface, timestamp of last used, expiration time, metric type and metric, the route state.

We provide an example in Figure 15 to comprehensively illustrate the route discovery process of DYMO in MANET. A path must be established to allow routing data from A to F. Similar to AODV, routing information is maintained at A in its own routing table while RREQ is flooded to A's neighbors and their other neighbors until the destination is reached. The sequence number in each RREQ is increased at each new neighbor before the RREQ is forwarded. At C, there are redundant RREQs from both directions with the same sequence number. If the alternative metric was not pre-set, it does not matter which path will be chosen. In this case, let us assume that C records the route from B to C. C then carries on forwarding this RREQ to the next hops, D and F. F is already the target destination while D is not. F replies the request with its RREP while D records the route. The RREP from F will be transferred on the reverse path from F back to C, B and A, the source. The data transmission will be then followed after this process has finished.

Upon sending the RREQ, the originator node A has to wait for a RREQ WAIT TIME for the RREP reception. In case there is no RREP received, A properly attempts to send another RREQ with a different sequence number. The waiting time is set to be 1000 milliseconds by default. Once the RREP is created and sent back, the flow of traffic of RREP is unicast. Therefore, it relies heavily on the MAC to ensure the bidirectional link underneath the network layer.

An example scenario of a CRN without PR is depicted in Figure 15. Since DYMO allows us to implement an alternative metric for the route selection process, we can modify the

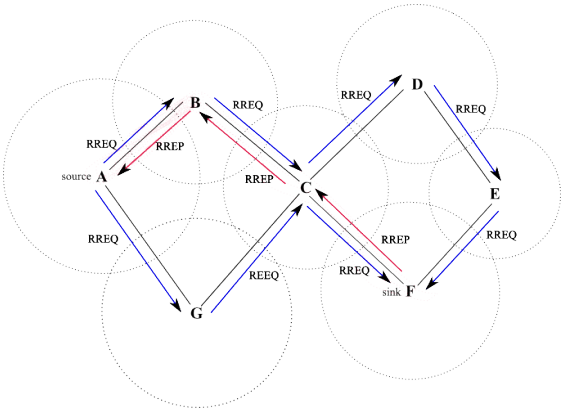


Figure 15. DYMO Route Discovery from source A to sink F without PR's existence

protocol with a minor change in retrieving MAC information according to the environment observations explained previously. An example of the application that DYMO can work in CRN is illustrated in Figure 16, in which we assume that B is kept in silent as it operates in the area where it overlaps with a PR transmitter. This is an expected route discovery with a scenario of a CRN with PR's existence depicted in Figure 16 using an alternative routing metric interference level.

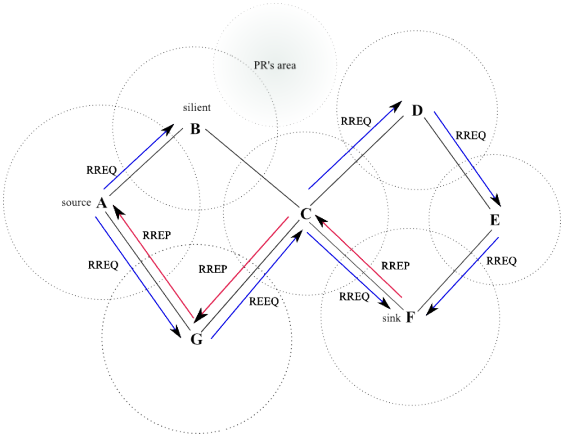


Figure 16. DYMO Route Discovery from source A to sink F with PR's existence

Route Maintenance: Each node that has received a route request and a route reply has to maintain the active route locally by monitoring the communication exchanging along the path. However, when there is problem in the network that triggers the path changes, a route error (RRER) may be generated and issued by these nodes to their neighbors. Two common problems are broken links on the current active route and unknown path from a received data packet.

Before using a route to forward a packet, a node must check the status of the route as follows. If the route is marked as broken, this node will not use it for forwarding. If the $Current\_Time$ is higher than the $Route.ExpirationTime$, the route has expired and cannot be used for forwarding. If the route is currently not in use when $(Current\_Time - Route.LastUsed) < (MAX\_SEQNUM\_LIFETIME)$,

the route table entry must be expunged. When generating a RRER, a node creates a list of unreachable destinations including their addresses with associated sequence number. Its routing table is also updated according to the broken link.

In the case of undeliverable packets, the RRER may be multicasting or unicasting to the neighbor from which the data packet was sent from. For unicast RRER, a special message type of value (TLV) is mandatorily included. In the case of broken link, the RRER is sent to all neighbors of the node that experience a broken link. After the other intermediate nodes receive the RRER, they verify the information contained in the RRER to react with the event accordingly. We illustrate this process in the example in Figure 17. A Route Error RRER is generated from C and then sent towards all of C's neighbors including B, G, and D to prevent future data transmission towards link C-F. These nodes had already recorded a path containing F; therefore, the RRER is further sent to A and E.

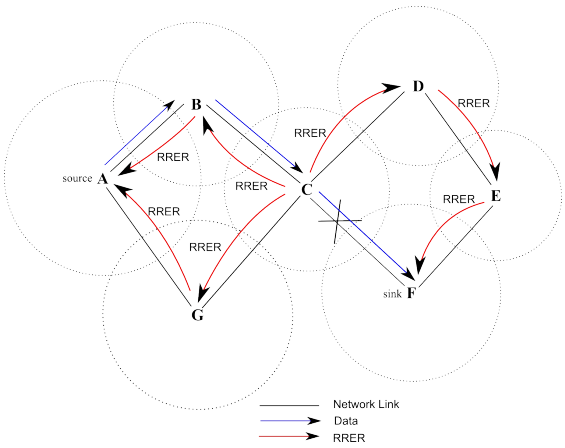


Figure 17. DYMO Route error sent from C when there is a broken link

The Hop Count is traditionally used as the default metric to select a route for a target node in exisiting table entries. However, some applications may require metric information other than Hop Count and this metric unfortunately sometimes causes the selection of the worst possible route in many situations. In CRNs, sometimes the Hop Count metric is not sufficient to choose a route to a target without interfering with current operating primary networks. The authors of AODVv2 [48] have been discussing alternative metrics since 2012. It was introduced in Internet draft version 24 [49]. Each such alternate metric measures a "cost" of using the associated route, and there are many different kinds of cost (latency, delay, monetary, energy, etc.).

The cost of each of the multiple routes is measured by a different metric. The specification provides the abstract function to evaluate the cost of each route in term of a function $Cost(R)$, where R is the route for which the cost is calculated. The route information for R must always include the type of metric by which $Cost(R)$ is evaluated. While using this alternative metric, we should also be careful and guarantee a loop-free environment while the routing engine is operating. In other words, given that the $Cost(R)$ is calculated, a loop-free routine should also be invoked. Since DYMO-AODVv2 is still work in progress, specific descriptions for metric type for alternative metric vaguely describes. The protocol requires experiments and changes if necessary. In the scope of this work, we tend to use it to evaluate the efficiency of interference level in preventing impact on PR networks while the routing engine is working.

B. CRN-DYMO routing protocol

Generally, our working scope is a conceptual framework that is illustrated in Figure 18. This framework focuses on an interference avoidance strategy which co-exists between the CRN and primary network scenarios, providing that the *Management Controller* stores all environments' observation parameters such as mobile nodes' movements and locations. The *Overlap Calculation* block then can locally retrieve, compute and analyse the current situation of the current CR compared to a PR emitter. Similarly, the *Node Density Estimator* obtains historical statistics of the node within a specific area and estimates the probability of current node density. These blocks produce estimated inputs for the *Interference Estimator*, which combines these inputs using fuzzy logic to compute the interference level.

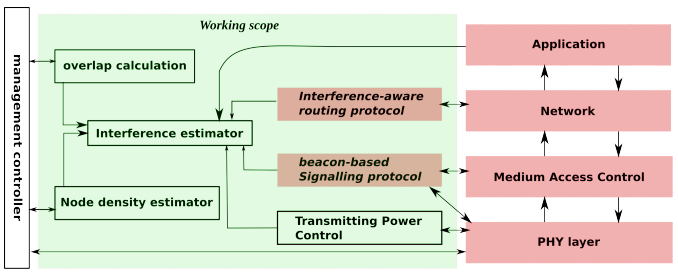


Figure 18. Working scope

Essentially, the DYMO routing agent can retrieve its neighbor's information based on the MAC layer and the interference level from the *Interference Level Estimator*. Customized DYMO-CRN the attaches interference level as an alternative metrics that is used for route selection, as explained in Figure 19 and Figure 20.

In Figure 19, we describe the route discovery, which is performed at node A to look for a path to node E overlapping with a PR. We assume that at C, the sequence numbers in the RREQs are the same for both direction from B to C and from G to C, but the interference levels from B and D are medium with non-zero probability. Therefore, in this case, the selected path to E is the path towards F. Figure 20 has the similar scenario with Figure 19 with an overlap with PR2 on G's and F's side. Route selection is changed since the interference levels from B and D are medium with non-zero probability while the interference levels from G and F are high with non-zero probability, and the interference level from C with PR2 is small with non-zero probability.

C. The routing process of Dymo-CRN routing protocol

We propose an extended Dymo routing protocol for CRN in this section. The proposed protocol involves route discovery and route maintenance processes.

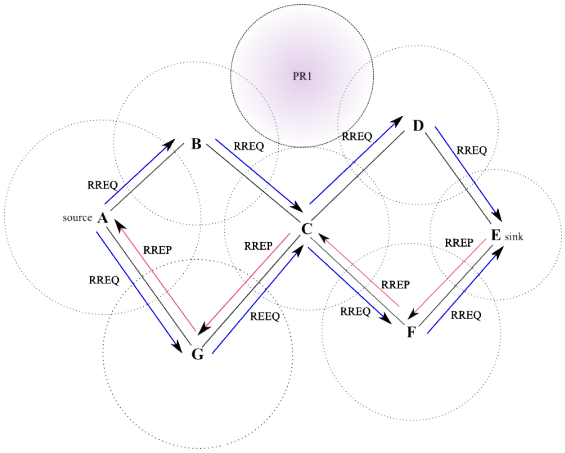


Figure 19. DYMO Route Discovery from source A to sink E with PR's existence

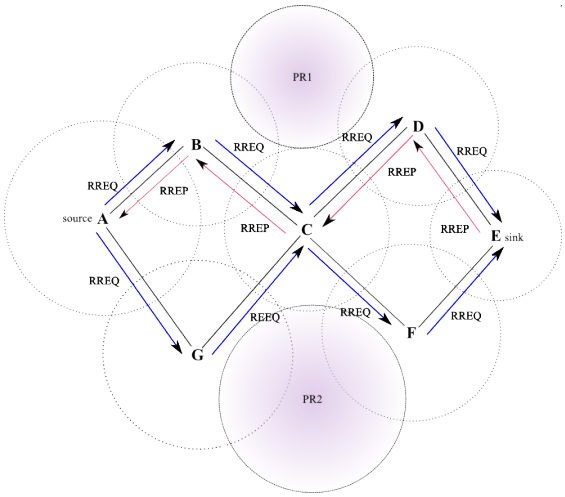


Figure 20. DYMO Route Discovery from source A to sink E with 2 PRs' existence

1) Route discovery: Similarly to Dymo specification, route discovery of the proposed protocol is performed when a node must transmit a packet towards a destination for which it does not have any route. The discovery process initiates and multicasts a route request message (RREQ) to find a path toward some target destination. Along the path, each of the neighbor who receives the request records a route toward the originator. When the request reaches the target node, this node also records the path included in the request and generates a unicast route reply (RREP) to the originator. Again, each neighbor along the path receiving this reply also records the path toward the target destination of the previous RREQ. These intermediate nodes then send the RREP unicast to the originator. The citeria to evaluate incoming route information are the hop count and the interference level (IL), which is the alternative metric. When a RREQ is received, each node will check, acquire its local interference level on each channel assuming that the node may have more than one available accessible channel. A comparison among its local IL is carried out and the most optimal IL is chosen as the local $metric\_IL$. As well as hop_count, the new_metric and new_metric_IL are recalculated as in (6) and (7). As presented, IL is a impact level computed from the percentage of overlap and node density degree. The new metric hence takes the average IL of the current local IL and the received IL. This is the metric of IL for the path from the orignator to the current node.

$$new\_metric\_IL = (metric\_IL + last\_metric\_IL)/2 \quad (6)$$

$$new\_metric = metric + last\_hop\_metric \quad (7)$$

The information from a received RREQ is handled by a routine call routing handling process (RteHandler), as described in Figure 21. By default, DYMO uses hop count as the routing metric to handle a route request with metric type 3. In the context of this work, we define metric type 4 to be the metric type of interference level. A metric type message TLV should hence be built and attached to a route message accordingly. During the RREQ process, the handler checks and looks in its routing table for an entry with the extracted metric type, e.g., metric type 4 for IL. In addition, an optimal IL is also acquired locally at the current node to recompute the new metric cost along the path. If the path to the target node exists, the sequence number of the entry is then compared with the sequence number in the received RREQ. In the end, the process will return an existing route entry without any update if the sequence number of the RREQ is outdated or the metric in the RREQ does not improve the cost of the path. If the routing entry needs to be updated (e.g., the sequence number needs to be refreshed, status of the route and update the new metric value). The routing table is also updated accordingly. The merit of enabling the metric IL is that we can guarantee the optimal IL from each CR to the its surrounding. If the current path has the minimum IL along the path, the route from the original CR to the target CR should have the least impact on any PR receivers.

2) Route maintenance: Route maintenance is performed to avoid permaturely expunging a route from the current route table as well as to avoid dropping packets when an active route breaks. Basically, the maintenance process consists of two operations, extending the route lifetimes upon successfully forwarding a packet and notifying the upstream nodes when a route to a target is broken due to loss of link to neighbors. In CRNs, breaking link to neighbors could be due to environmental change such as the transmission of PR receivers. However, thanks to the information obtained from the radio event table, CR nodes can always look for an alternative route and start notifying their neighbors about the change. This change would result in a complete path change (via different channels). The discovery process may be invoked in the case of one of the nodes is unreachable. This work can be deferred to future work.

VI. Conclusion

Overall, we have presented the entirely of a routing strategy in CRNs. The article covers everything from the simplest observation of the surroundings regarding the overlap phenomena and primary node density with associated techniques. These observations are inputs to produce a routing metric that could prevent foreseen interference when a CRN coexists

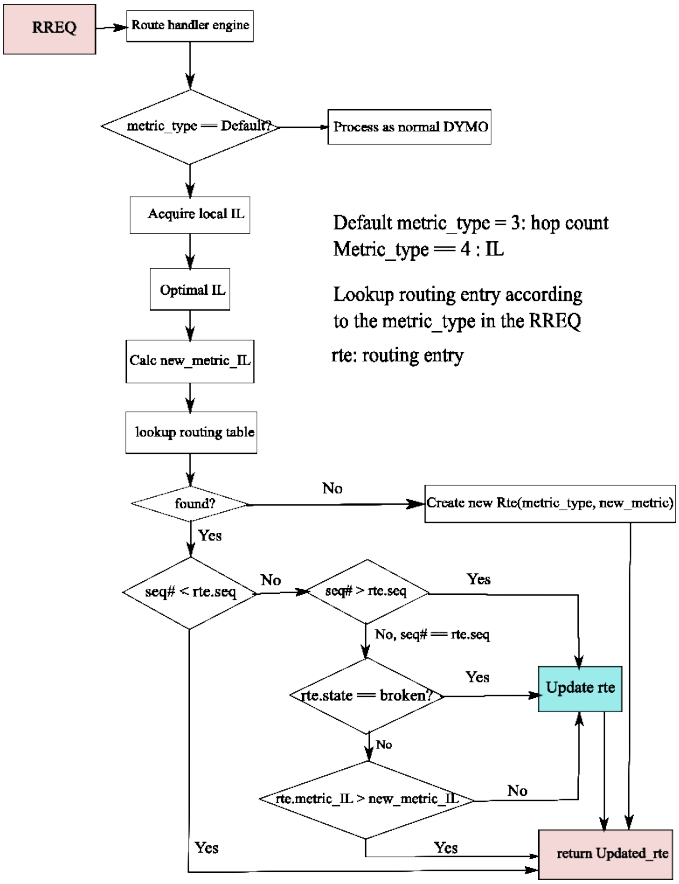


Figure 21. Routing information process

with a primary network. Essentially, we introduced a medium access control messages dissemination procedure, in which CR nodes maintain the knowledge of their neighbors as well as their resources before they can initiate transmissions. This establishment supports the routing protocol at the upper layer, which in this case is the interference level routing protocol. This is a guideline for developing a cognitive routing protocol that could prevent potential impact on primary networks when CRNs operate. Further extension of this work is encouraged to provide a more elaborate protocol so that it can be applied in a real CRN platform.

ACKNOWLEDGMENT

This work is supported by LICoRNe project, funded in part by the National Agency for Research in France - ANR (Agence Nationale de la Recherche). Besides, the authors would like to thank Dr. Yao-ban Chan of University of Queensland for proofreading.

REFERENCES

[1] M. T. Quach, F. Krief, M. A. Chalouf, and H. Khalifé, "Fuzzy-based interference level estimation in cognitive radio networks," in The Tenth Advanced International Conference on Telecommunications (AICT). IARIA XPS, 2014, pp. 138–143.

[2] M. J. Copps, "Bringing broadband to rural america: Report on a rural broadband strategy," Federal Communications Comission, Tech. Rep., May 2009.

[3] F. C. Comission, "Bringing broadband to rural america: Update to report on a rural broadband strategy," Federal Communications Commission, Tech. Rep., 2011.

[4] J. Genachowski, M. Clyburn, and J. Rosenworcel, "Eight broadband progress report," Federal Communications Commission, Tech. Rep. FCC 12-90, August 2012.

[5] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dymanic spectrum access/cognitive radio wireless newtork: A survey," in Computer Networks, vol. 50, 2006, pp. 2127–2159.

[6] D. N. Ekram Hossain and Z. Han, Dynamic spectrum access and management in cognitive radio networks. Cambridge University Press Cambridge, 2009.

[7] T. J. Ross, Fuzzy logic with engineering applications. John Wiley & Sons, Ltd., 2010.

[8] M. T. Quach and H. Khalife, "The impact of overlap regions in cognitive radio networks," in Wireless Days (WD), IFIP, 2012, pp. 1–3.

[9] M. T. Quach, D. Ouattara, F. Krief, H. Khalifé, and M. A. Chalouf, "Overlap regions and grey model-based approach for interference avoidance in cognitive radio networks," in IEEE Fifth International Conference on Ubiquitous and Future Networks (ICUFN), 2013, pp. 642–647.

[10] M. Cesana, F. Cuomo, and E. Ekici, "Routing in cognitive radio networks: Challenges and solutions," Ad Hoc Networks, vol. 9, no. 3, 2011, pp. 228–248.

[11] B. Wang and K. J. R. Liu, "Advances in cognitive radio networks: A survey," IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 1, Feb 2011, pp. 5–23.

[12] Q. Wang and H. Zeng, "Route and spectrum selection in dynamic spectrum networks," in IEEE Consumer Communications and Networking Conference (CNCC), 2006, pp. 342–346.

[13] H. Khalifé, N. Malouch, and S. Fdida, "Multihop cognitive radio networks: To route or not to route," IEEE NET, The Magazine of Global Internetworking, vol. 23, no. 4, August 2009, pp. 20–25.

[14] G. Cheng, W. Liu, Y. Li, and W. Cheng, "Spectrum aware on-demand routing in cognitive radio networks," in 2nd International Symposium on New Frontiers in Dynamic Spectrum Access Networks. IEEE, 2007, pp. 571–574.

[15] Y. Liu, L. Cai, and X. Shen, "Spectrum-aware opportunistic routing in multi-hop cognitive radio networks," IEEE Journal on Selected Areas in Communications, vol. 30, no. 10, November 2012, pp. 1958–1968.

[16] G.-M. Zhu, I. F. Akyildiz, and G.-S. Kuo, "Stod-rp: A spectrum-tree based on-demand routing protocol for multi-hop cognitive radio networks," in IEEE Global Telecommunications Conference, Nov 2008, pp. 1–5.

[17] A. Cacciapuoti, C. Calcagno, M. Caleffi, and L. Paura, "CAODV: Routing in mobile ad-hoc cognitive radio networks," in Wireless Days (WD), October 2010, pp. 1–5.

[18] S. Salim and S. Moh, "On-demand routing protocols for cognitive radio ad hoc networks," EURASIP Journal on Wireless Communications and Networking, vol. 2013, no. 1, 2013, pp. 1–10.

[19] C. Xin, B. Xie, and C.-C. Shen, "A novel layered graph model for topology formation and routing in dynamic spectrum access networks," in First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks. IEEE, 2005, pp. 308–317.

[20] S. Krishnamurthy, M. Thoppian, S. Venkatesan, and R. Prakash, "Control channel based mac-layer configuration, routing and situation awareness for cognitive radio networks," in Military Communications Conference (MILCOM). IEEE, 2005, pp. 455–460.

[21] Q. Guan, F. Yu, S. Jiang, and G. Wei, "Prediction-based topology control and routing in cognitive radio mobile ad hoc networks," IEEE Transactions on Vehicular Technology, vol. 59, no. 9, Nov 2010, pp. 4443–4452.

[22] K. Chowdhury and M. Felice, "Search: A routing protocol for mobile cognitive radio ad-hoc networks," Computer Communications, vol. 32, no. 18, 2009, pp. 1983–1997.

[23] K. Habak, M. Abdelatif, H. Hagrass, K. Rizc, and M. Youssef, "A location-aided routing protocol for cognitive radio networks," in International Conference on Computing, Networking and Communications (ICNC), Jan 2013, pp. 729–733.

[24] K. Chowdhury and I. F. Akyildiz, "Crp: A routing protocol for cognitive radio ad hoc networks," Selected Areas in Communications, IEEE Journal on, vol. 29, no. 4, 2011, pp. 794–804.

[25] E. Hossain, L. Le, N. Devroye, and M. Vu, "Cognitive radio: From theory to practical network engineering," in New Directions in Wireless Communications Research, V. Tarokh, Ed. Springer, 2009, pp. 251–289.

[26] T. C. Clancy, "Achievable capacity under the interference temperature model," in 26th IEEE International Conference on Computer Communications (INFOCOM), May 2007, pp. 794–802.

[27] M. Gastpar, "On capacity under receive and spatial spectrum-sharing constraints," IEEE Transactions on Information Theory, vol. 53, no. 2, Feb 2007, pp. 471–487.

[28] P. J. Kolodzy, "Interference temperature: a metric for dynamic spectrum utilization," International Journal of Network Management, vol. 16, no. 2, 2006, pp. 103–113.

[29] W. Wang, T. Peng, and W. Wang, "Optimal power control under interference temperature constraints in cognitive radio network," in IEEE Wireless Communications and Networking Conference, WCNC. IEEE, 2007, pp. 116–120.

[30] Y. Xing, C. Mathur, M. Haleem, R. Chandramouli, and K. Subbalakshmi, "Dynamic spectrum access with qos and interference temperature constraints," IEEE Transactions on Mobile Computing, vol. 6, no. 4, April 2007, pp. 423–433.

[31] M. Vu, S. Ghassemzadeh, and V. Tarokh, "Interference in a cognitive network with beacon," in IEEE Wireless Communications and Networking Conference (WCNC), March 2008, pp. 876–881.

[32] D. Julog, "The basis of grey theory," Huazhong University of Science and Technology Press, 2002.

[33] L. Sifeng, F. J., and Y. Yingjie, "A brief introduction to grey systems theory," in Grey Systems and Intelligent Services (GSIS), 2011 IEEE International Conference on, 2011, pp. 1–9.

[34] E. Kayacan, B. Ulutas, and O. Kaynak, "Grey system theory-based models in time series prediction," Expert Systems with Applications, vol. 37, no. 2, 2010, pp. 1784–1789.

[35] D. Ouattara, F. Krief, M. A. Chalouf, and O. Ahmdi, "Spectrum sensing improvement in cognitive radio networks for real-time patients monitoring," in International Conference on Wireless Mobile Communication and Healthcare (MobiHealth), November 2012, pp. 179–188.

[36] E. Hossain, D. Niyato, and Z. Han, "Dynamic spectrum access and management, cognitive radio networks," Cambridge University Press, 2009.

[37] D. B. Johnson and D. A. Maltz, Dynamic source routing in ad hoc wireless networks. Springer, 1996, ch. 5, pp. 153–181.

[38] D. Johnson, Y. Hu, and D. Maltz, "The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4," RFC 4728 (Experimental), Internet Engineering Task Force, Feb. 2007, URL: http://www.ietf.org/rfc/rfc4728.txt [accessed: 2013-11-02].

[39] C. E. Perkins, E. Belding-Royer, and E. M. Royer, "Ad-hoc on-demand distance vector routing," in Second IEEE Workshop on Mobile Computing Systems and Applications. IEEE, 1999, pp. 90–100.

[40] C. E. Perkins, E. Belding-Royer, and S. R. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," RFC 3561 (Experimental), Internet Engineering Task Force, 2003, URL: http://www.ietf.org/rfc/rfc3561.txt [accessed: 2013-11-02].

[41] C. E. Perkins, E. M. Royer, S. R. Das, and M. K. Marina, "Performance comparison of two on-demand routing protocols for ad hoc networks," Personal Communications, IEEE, vol. 8, no. 1, 2001, pp. 16–28.

[42] M. Jain, J. I. Choi, T. Kim, DineshBharadia, S. Seth, K. Srinivasan, P. Levis, S. Katti, and P. Sinha, "Practical, real-time, full duplex wireless," in Proceedings of the 17th annual international conference on Mobile computing and networking. ACM, 2011, pp. 301–312.

[43] P. Manickam, T. G. Baskar, M. Girija, and D. Manimegalai, "Performance comparisons of routing protocols in mobile ad hoc networks," International Journal of Wireless & Mobile Networks, vol. 3, no. 1, 2011, pp. 98–106.

[44] S. Sagar, J. Saqib, A. Bibi, and N. Javaid, "Evaluating and comparing the performance of dymo and olsr in manets and in vanets," in IEEE 14th International Multitopic Conference (INMIC), 2011, pp. 362–366.

[45] C. Perkins and I. C. Futurewei, "Dynamic manet on-demand (AODVv2) routing," Internet-Draft, 2013, URL: https://tools.ietf.org/html/draft-ietf-manet-dymo-26 [accessed: 2013-11-15].

[46] N. Sivakumar and S. K. Jaiswal, "Comparison of dymo protocol with respect to various quantitative performance metrics," Department of Computer Science, Malardalen University, 2009.

[47] S. K. Bisoyi and S. Sahu, "Performance analysis of dynamic manet on-demand (dymo) routing protocol," Special Issue of IJCCT, vol. 1, no. 2, 2010, p. 3.

[48] C. Perkins, Futurewei, S. Ratliff, Cisco, J. Dowdell, and Cassidian, "Dynamic manet on-demand (AODVv2) routing, draft-ietf-manet-aodvv2-04," Internet-Draft, 2006, URL: https://tools.ietf.org/html/draft-ietf-manet-dymo-04 [accessed: 2013-11-14].

[49] C. Perkins and I. C. Futurewei, "Dynamic manet on-demand (AODVv2) routing," Internet-Draft, 2012, URL: https://tools.ietf.org/html/draft-ietf-manet-dymo-24 [accessed: 2013-11-15].

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601