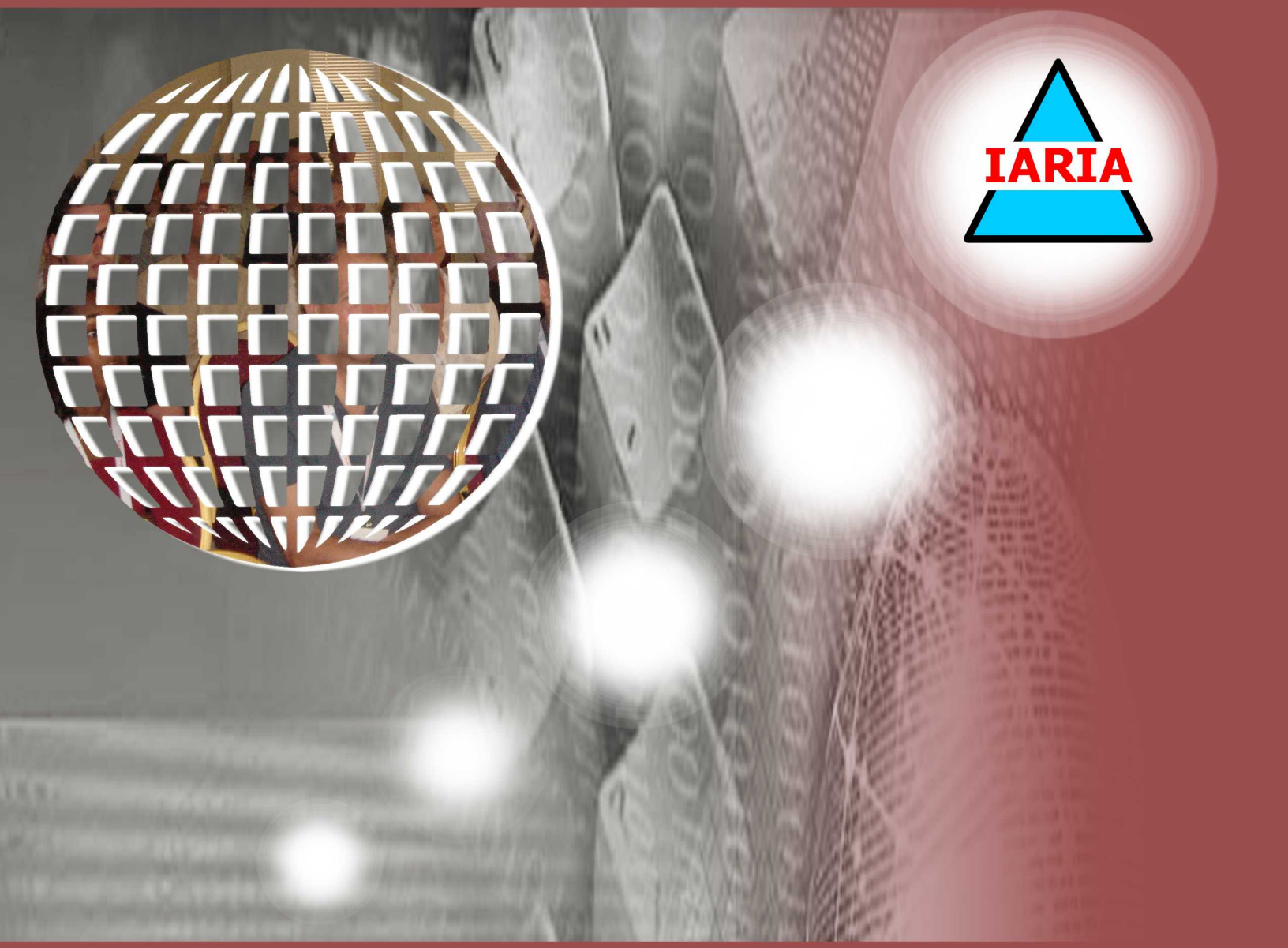
International Journal on Advances in Telecommunications



2012 vol. 5 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 5, Number 1 & 2, 2012

Javier Barria, Imperial College London, UK
Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Mark Bentum, University of Twente, The Netherlands
David Bernstein, Huawei Technologies, Ltd., USA
Eugen Borgoci, University "Politehnica"of Bucharest (UPB), Romania
Christos Bouras, University of Patras, Greece
David Boyle, Tyndall National Institute, University College Cork, Ireland
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, University of Sao Paulo, Brazil
Hakima Chaouchi, Telecom SudParis, France
Emmanuel Chaput, IRIT-CNRS, France
Silviu Ciochina, Universitatea Politehnica din Bucuresti, Romania
José Coimbra, Universidade do Algarve, Portugal
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Denis Collange, Orange Labs, France
Noel Crespi, Institut TELECOM SudParis-Evry, France
Leonardo Dagui de Oliveira, Escola Politécnica da Universidade de São Paulo, Brazil
Gerard Damm, Alcatel-Lucent, USA
Danco Davcev, University for Information Science & Technology "St. Paul the Apostle", Ohrid, Macedonia
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
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
Haibin Liu, China Aerospace Science and Technology Corporation, 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
Reijo Savola, VTT Technical Research Centre of Finland, Finland
Marialisa Scatà, University of Catania, Italy
Zary Segall, Chair Professor, Royal Institute of Technology, Sweden
Sergei Semenov, Renesas Mobile Europe, 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
Danai Skournetou, Tampere University of Technology, Finland
Mariusz Skrocki, Orange Labs Poland / Telekomunikacja Polska S.A., Poland

Anita Sobe, Klagenfurt University, Austria
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica Spahiu, University of Craiova, Romania
Radu Stoleru, Texas A&M University, USA
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 5, Numbers 1 & 2, 2012

CONTENTS

Adaptive Jitter Buffer based on Quality Optimization under Bursty Packet Loss

Liyun Pang
Audio Technology Team
Huawei European Research Center
Munich, Germany
liyun.pang@huawei.com

Laszlo Böszörmenyi
Department of Information Technology
University Klagenfurt
Klagenfurt, Austria
office-lb@itec.uni-klu.ac.at

***Abstract*—Quality of voice delivered over packet networks is affected by various factors such as packet loss, end-to-end delay, packet delay variation (jitter) and codec bit rate. Different approaches and models predict speech quality as a function of such impairments. In order to ensure a continuous play-out of voice transmitted over a packet switched network, jitter buffers are commonly used to counter jitter introduced by queuing. In this paper, we propose a new adaptive jitter buffer algorithm based on optimizing the predicted voice quality. The algorithm consists of an adaptive play-out mechanism based on the extended E-Model taking into account packet loss pattern and a time-scaling technique relying on a speech classification mechanism embedded in the decoder. In our work, we apply the time-scaling to the modified AMR WB decoder. Simulation results show that the proposed algorithm outperforms the best existing algorithms in a variety of different network scenarios under bursty packet loss.**

Keywords - adaptive jitter buffer; E-Model; AMR WB; time-scaling; bursty packet loss

I. INTRODUCTION

Transport of Voice over IP (VoIP) is one of the most important applications among recent IP-based telecommunication services. VoIP can be seen as an alternative to the traditional circuit switched telephony with the advantages of reduced cost, simplified network and simplified network management. The main challenge is to guarantee the same Quality of Service (QoS) as that of traditional telephony. Voice quality is the key metric for QoS for VoIP applications. Packet losses, latency (delay) and delay variation (jitter) are the major factors, inevitable in a packet network, contributing to speech quality degradation. In particular, the delay jitter introduced by queuing in packet switched networks has a devastating impact on the perceived quality. In order to smooth delay jitter, a jitter buffer mechanism is required at the receiver for ensuring a continuous play-out of voice data.

When a jitter buffer is applied, received packets are stored in the buffer after arrival, and played out sequentially at scheduled times. Any packet arriving after its scheduled playout time is discarded at the receiver, resulting in the so-called late loss. The late loss rate can be reduced by scheduling a later playout time at the expense of an excessively longer end-to-end delay. The problem of delay jitter is thereby converted into end-to-end delay and packet loss. Previous work mainly focused on designing jitter buffers solely based on the trade-off between end-to-end delay (playout delay) and packet loss rate due to late arrival. The playout delay is adjusted either at the beginning of each talk-spurt [2][3] (called per-talk-spurt), or more adaptively within the speech talk-spurt using time-scale modification [4][5] (called per-packet). Although such designs can achieve a minimum average end-to-end delay for a specified packet loss rate, they do not take into account the overall perceived speech quality. Recently, much effort has been devoted to developing approaches to adjust the jitter buffer with the objective of optimizing the perceived speech quality given by the Mean Opinion Score (MOS) [1][6][7][8]. To develop such quality-based approaches, the non-intrusive parametric models which estimate MOS values directly from network parameters and terminal characteristics are required. The ITU-T E-Model [9] is one of the most well-known parametric models. The output of the E-Model, the so called *R* value, can be easily mapped onto a corresponding MOS value using a transformation given in Appendix I of [9].

The ITU-T E-Model has been initially developed as a network planning tool [10]. Although the E-Model has limited accuracy for evaluating conversational speech quality [11][12][13][14], it has shown applicability in the context of QoS monitoring [15][16][17]. In [18][19][20], quality-based playout scheduling approaches were proposed to maximize perceived speech quality using the *R* value of the E-Model as cost function. These approaches adjust the playout delay on per-talk-spurt basis and their performance is limited when talk-spurts are long and the network delay varies significantly. A per-packet quality-based jitter buffer algorithm is described in [21]. The playout delay estimation is designed as an unconstrained optimization problem that maximizes the *R* value. However, in per-packet jitter buffer management, a speech frame can only be time-scaled within a certain range to maintain the naturalness of the original speech signal [4]. Thus, a constrained optimization problem is more suitable.

To design a quality-based jitter buffer algorithm, an estimate of network delay distribution is required. Some works assume a certain parametric model to estimate the Cumulative Distribution Function (CDF) of the network delay distribution, such as Pareto [21], Weibull [18] and Gamma [22]. In fact, delay and jitter in a VoIP session are non-stationary and have a high degree of variability even within a single session. In particular for jitter buffer

management on per-packet basis, the network delay behavior cannot be modeled just by a certain type of distribution.

Many previous works focus on the packet loss impairment [11][12][13][23][24]. Some studies assume random packet loss for quality estimation and focus only on the overall packet loss rate [18][25]. Several studies revealed temporal dependencies in packet loss based on network statistics. The overall loss rate alone is not sufficient to predict the speech quality perceived by users [26]. The authors proposed a method to calculate the burstiness level from the packet loss pattern which can be converted into an equivalent random packet loss factor [27]. The results in [28] have shown that other properties such as the loss location and the loss distribution also have impact on the perceived speech quality.

In [1], we presented an adaptive jitter buffer system implementing per-packet scheduling based on the extended E-Model. The system contains a spike detection mechanism and a classifier based time-scaling technique similar to that proposed in [5]. The time-scaling technique is implemented directly inside the *AMR-NB* decoder [29]. It is advantageous for the quality and makes it possible to use the internal parameters such as the pitch lag and the gains for time-scaling. In this paper, we extend our previous work by taking into account bursty packet loss and adapting the mechanism to wideband speech transmission for *AMR-WB* [30][31].

The rest of the paper is organized as follows. Section II gives a brief overview of the extended E-Models including the impairment model for random packet loss and bursty packet loss. In Section III, the play-out algorithm based on optimizing the predicted voice quality is proposed. Section IV presents the modified time-scaling embedded in the *AMR-WB* decoder. Simulation results illustrating the performance of the proposed scheme are presented in Section V. Finally, we conclude this paper in Section VI.

II. EXTENDED E-MODEL

A. ITU-T E-Model

The ITU-T E-Model is a computational model for the prediction of the expected voice quality which combines different impairments contributing to speech quality degradation, such as loudness, background noise, low bit-rate coding distortion, codec, echo, packet loss and delay. A large set of these impairment factors have been quantified regarding their impact on the conversational speech quality. The underlying assumption of the E-Model is that all those impairment factors are additive on a psychological scale, and summed to form a rating factor *R*. The rating factor lies in the range of 0 to 100. An invertible mapping exists between *R* and conversational MOS. A rating of '0' represents a MOS value '1' (bad quality) and '100' of *R* represents MOS value '4.5' (high quality). For wideband speech transmission, *R* can go beyond 100. The output *R* value is obtained by subtracting impairment factors from a basic quality measure [9]:

$$R = R_0 - I_s - I_d - I_{e,eff} + A \quad (1)$$

where R_0 represents the basic signal-to-noise ratio; I_s is the Simultaneous Impairment Factor which occurs more or less simultaneously with the speech signal; I_d represents the impairments caused by delay; $I_{e,eff}$ is the Effective Equipment Impairment Factor representing impairments caused by low bit rate codecs and packet loss. A is an Advantage Factor which has accordingly no relationship to any other parameter and normally can be neglected. $I_{e,eff}$ and I_d are the most important factors to predict voice quality in packet networks.

B. Effective Equipment Impairment Factor

The model of $I_{e,eff}$ in [18] is defined as

$$I_{e,eff} = I_e + (95 - I_e) \cdot \frac{P_{pl}}{\frac{P_{pl}}{BurstR} + B_{pl}} \quad (2)$$

where I_e is a codec specific value and represents the equipment factor in loss-free networks. B_{pl} represents the robustness of the codec in lossy networks. ITU-T G.113 Appendix [32] lists the provisional reference values of several codecs, derived from subjective MOS test results and network experience. Both the packet loss probability P_{pl} and the Burst Ratio $BurstR$ depend on the packet loss pattern. If $BurstR$ = 1, the packet loss is random. If $BurstR$ >1, the packet loss is bursty. Otherwise, the packet loss distribution is less bursty. Bursty packet loss can be emulated by a two state Markov model characterized by two transition probabilities: p between "Found" and "Loss" state, and q between "Loss" and "Found" state. p and q can be obtained from the mean loss rate (mlr) and the mean burst length (mbl) as

$$p = \frac{mlr}{mbl(1 - mlr)}, q = \frac{1}{mbl} \quad (3)$$

$BurstR$ can be calculated as [18]

$$BurstR = \frac{1}{p + q} \quad (4)$$

By combining (3) and (4), $BurstR$ is calculated as

$$BurstR = mbl(1 - mlr) \quad (5)$$

As $P_{pl} = mlr \cdot 100$, we can express $BurstR$ as a function of mbl and P_{pl}

$$BurstR = \left(1 - \frac{P_{pl}}{100}\right) \cdot mbl \quad (6)$$

As mbl represents the average length of burst in an arrival sequence, it can be calculated by counting the number of consecutively lost packets and multiplying the count with the corresponding probability as

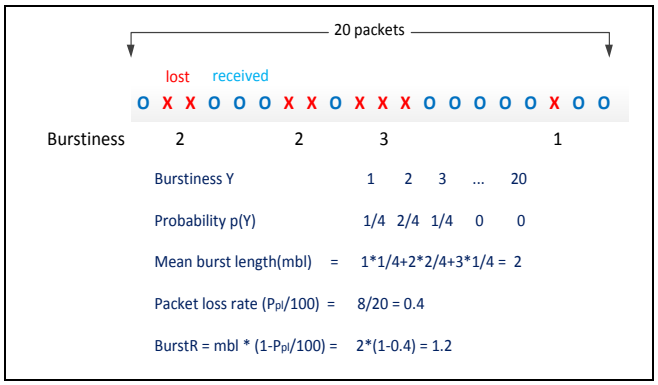


Figure 1. Example of *BurstR* Calculation

TABLE I. PROVISIONAL VALUE

Codec	I_e	B_{pl}
AMR-WB 12.65 kbit/s	20	4.3
AMR-NB 12.2 kbit/s	5	10

$$mbl = \sum_{k=0}^{\infty} kP(Y=k) \quad (7)$$

where $P(Y=k)$ is the probability for having k consecutively lost packets. With (6) and (7), we can obtain $BurstR$ as

$$BurstR = \left(1-\frac{P_{pl}}{100}\right)\sum_{k=0}^{\infty} kP(Y=k) \quad (8)$$

Fig. 1 gives an example of how to calculate the $BurstR$ for a particular sequence of packets. The overall number of packets is assumed to be 20. The packet loss pattern is described by color in Fig. 1: the red X means a lost frame and the blue O represents a successfully received frame. By giving the sequence of error information, we can calculate an instantaneous $BurstR$ as 1.2 for this sequence, as shown in Fig. 1.

We further divide P_{pl} into two parts

$$P_{pl} = 100 \cdot (p_n + p_b) \quad (9)$$

where p_n is the packet loss rate in the network and p_b is the late packet loss rate caused by packet drops in the jitter buffer. Since a packet is discarded when it arrives after its scheduled playout time, the late loss rate p_b is calculated as

$$\begin{aligned} p_b &= (1-p_n)\big(1-P_r(X \le d)\big) \\ &= (1-p_n)\big(1-F(d)\big) \end{aligned} \quad (10)$$

with $F(d)$ being the CDF of network delay (d) obtained from histogram statistics of previous network delays.

The provisional planning values of I_e and B_{pl} required in (2) can be found in [32][33], and the values of *AMR-NB/AMR-WB* are listed in Table I. Applying these values to (2), we can calculate the $I_{e,eff}$ for *AMR-NB* 12.2 kbit/s as

$$\begin{aligned} I_{e,eff} &= 5 + 90\frac{P_{pl}}{\frac{P_{pl}}{BurstR}+10} \\ &= 5 + 90\frac{p_n+p_b}{\frac{p_n+p_b}{BurstR}+\frac{10}{100}} \end{aligned} \quad (11)$$

However, the provisional reference values have to be derived from a large quantity of subjective tests. In order to avoid subjective tests, alternatively, $I_{e,eff}$ can be estimated with a logarithmic fitness curve as given in [23][24]:

$$I_{e,eff} = A + B\,ln(1 + C(p_n + p_b)) \quad (12)$$

where *A*, *B* and *C* are curve fitting parameters. Other codecs may have different forms of curve for $I_{e,eff}$ [34]. The empirical formula for *AMR-NB* 12.2 kbit/s is [18]

$$I_{e,eff} = 14.96 + 16.68\,ln(1 + 30.11(p_n + p_b)) \quad (13)$$

A similar equation to (2) for the wideband effective equipment factor $I_{e,eff,WB}$ under random packet loss ($BurstR = 1$) is given by [9][35][36]. Similarly, for wideband speech codec such as *AMR-WB* 12.65 kbit/s, $I_{e,eff,WB}$ can be obtained by applying provisioning values from [33] for random packet loss

$$\begin{aligned} I_{e,eff,WB} &= 20 + 75\frac{P_{pl}}{P_{pl}+4.3} \\ &= 20 + 75\frac{p_n+p_b}{p_n+p_b+\frac{4.3}{100}} \end{aligned} \quad (14)$$

For bursty packet loss, an empirical formula for $I_{e,eff,WB}$ was proposed in [13][26] using Genetic Programming for different wideband codecs. For example, the $I_{e,eff,WB}$ for *AMR-WB* 12.65 kbit/s is expressed as

$$\begin{aligned} I_{e,eff,WB} = \Bigg\{ &\ln\left(\frac{9\cdot(43.91+P_{pl}\cdot 187.62^2)}{mbl^5-P_{pl}}\right) + P_{pl} + 43.91 + \\ &187.62\cdot P_{pl}\Bigg\}\cdot 0.8303 + 8.9977 \end{aligned} \quad (15)$$

C. Delay Impairment Factor

If I_d refers to impairments only due to end-to-end delay d, then I_d can be derived by curve fitting as described in [25]

$$I_d = 0.024d + 0.11(d-177.3)H(d-177.3) \quad (16)$$

where $H(x)$ is the step function ($H(x)=0$ if $x<0$; $H(x)=1$ else).

D. Extended E-Model

For the extended E-Model, we assume that impairments due to other factors such as echo or delay are not present. All input parameters and their recommended ranges are found in [9]. For those parameters which are not available at the time of planning, the default values from the ITU [32][37] are recommended. We only focus on IP networks, and the expression of E-Model in (1) can be simplified in terms of transport-level metrics [25] for narrow band

$$R_{NB} = 93.2 - I_d - I_{e,eff} \quad (17)$$

And a similar expression for wide band [25] is

$$R_{WB} = 129 - I_d - I_{e,eff,WB} \quad (18)$$

If we define the sum of I_d and $I_{e,eff}$ as a new impairment factor I

$$I = I_d + I_{e,eff} \quad (19)$$

then both (17) and (18) can be simplified as

$$R = R_{max} - I \quad (20)$$

where R_{max} is 93.2 for narrow band and 129 for wide band.

This formulation of R in (20) is used as the cost function in our jitter buffer management to estimate the playout delay by maximizing R which is equivalent to minimizing I. Equation (11) with $BurstR$ = 1 is already used in [1] for *AMR-NB* random packet loss. We further investigate (11) for modeling $I_{e,eff}$ of *AMR-NB* under bursty packet loss, (14) for modeling $I_{e,eff,WB}$ of *AMR-WB* under random packet loss, and also the empirical formula (15) for modeling $I_{e,eff,WB}$ of *AMR-WB* under bursty packet loss. The results are shown in Section V.

III. Proposed Play-out Algorithm

The proposed receiver includes an adaptive jitter buffer algorithm and a time-scaling embedded inside the decoder, as shown in Fig. 2. The adaptive play-out algorithm is the main control unit. Since spikes are very common in VoIP transmission, spike detection in [3] is implemented to switch between NORMAL mode and SPIKE mode. In SPIKE mode, the scheduled playout time follows current network condition. In NORMAL mode, the scheduled playout time is estimated based on the extended E-Model, as discussed in Section II.

The play-out algorithm is similar to the one proposed in [1], and will be described using the same basic notations listed in Table II. For each packet, a certain playout time is scheduled at the receiver before its arrival. When a packet arrives at the receiver before its scheduled time, it can be played out without packet loss. Before playing out the current speech frame, the playout delay of the next expected packet has to be estimated to obtain the expected frame length of the current frame. The playout delay is chosen in order to maximize the predicted speech quality in terms of R. As discussed in Section II, R depends on the end-to-end delay d, network loss rate p_n, late loss rate p_b and also $BurstR$ regarding bursty packet loss. Both network loss rate and $BurstR$ can be calculated based on the loss pattern of previous received packets stored in a history window with the window size W. The late loss rate is determined by the playout buffering algorithm, and thus by the end-to-end delay (playout delay). Therefore, (20) can be expressed as a function of playout delay, and applied as the cost function in the playout buffering algorithm to predict the voice quality.

The playout delay for each packet is estimated based on maximizing the expected R value. The operation of the jitter buffer is based on the statistics of the delay and packet loss of the previous received packets.

The algorithm works as follows

1. Receive a new $packet^i$, and obtain network delay information d_n^i and error information from the RTP header. The loss pattern of the most recent received W (history window size) packets is updated and $BurstR$ is calculated (default is 1).
2. Spike detection: check the current network condition, and switch between SPIKE/ NORMAL.
3. Playout time scheduling
   a) If this is the first packet of the talk-spurt, follow network delay
      $d_p^i = d_n^i$
   b) Otherwise, use the estimated playout delay
      $d_p^i = \hat{d}_p^i$
4. Playout delay estimation
   a) SPIKE: follow the current network delay $\hat{d}_p^{i+1} = d_n^i$, and skip step 5.
   b) NORMAL: estimate playout delay based on the E-Model.
5. E-Model based playout delay estimation in NORMAL mode
   a) Update delay statistics of the most recent received W (history window size) packets only in NORMAL mode
   b) Find the optimal playout delay for $packet^{i+1}$
      $\hat{d}_p^{i+1}: I_m(\hat{d}_p^{i+1}) = \min_{d_{min} \le d \le d_{max}} I_m(d)$
      where d_{min} and d_{max} are the constraints specified by the time-scaling to make the artifacts less audible:
      $d_{min} = d_p^i - (L_o - L_{min})$
      $d_{max} = d_p^i + (L_{max} - L_o)$
6. Calculate the new length of packet^i
   $\Delta^i = \hat{d}_p^{i+1} - d_p^i$
   $L^i = L_o + \Delta^i$
7. Send $packet^i$ and expected length L^i to the decoder.

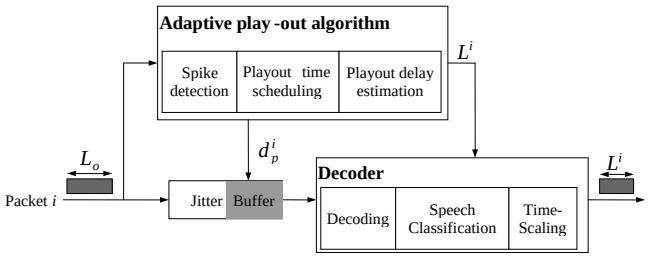


Figure 2. Proposed adaptive jitter buffer at the receiver

TABLE II. BASIC NOTATIONS

symbol	Definition
d_n^i	network delay of packet i
d_p^i	actual playout delay of packet i
$\hat{d}_p^i$	estimated playout delay of packet i
L_o	original frame length, 160 samples for AMR
L^i	modified frame length of packet i
Δ^i	frame length difference of packet i
L_{max}	maximum possible time-scaled frame length
L_{min}	minimum possible time-scaled frame length

IV. TIME-SCALING EMBEDDED IN THE DECODER

The E-Model based playout scheduling algorithm described in Section III is applied specifically to the CELP codec. The standard 3GPP *AMR-WB* decoder [30] is modified to embed the time-scaling technique based on speech classification. According to the evaluated frame type, different time-scaling (extension or suppression) operations are applied to the frame in the excitation domain.

A. Speech Classification

In our previous work [1], the speech classification was based on three parameters: the Voicing Factor, the Spectral Ratio and the Energy Variation. The frame type is thus identified by comparing them to the predefined reference values. Special frames such as plosive or over-voiced frames are also differentiated from others by using the internal parameters inside the *AMR-NB* decoder. In order to make the classification more accurate, we implement the merit function f_m which has been defined in the *VMR* codec [38] for the frame erasure concealment as

$$f_m = \frac{1}{7}\left(2\bar{R}_{xy}^s + e_{tilt}^s + SNR^s + pc^s + E_{rel}^s + zc^s\right) \qquad (21)$$

where $[\cdot]^s$ represents the scaled version of the corresponding classification parameters, including: the normalized correlations $\bar{R}_{xy}$, the spectral tilt parameter e_{tilt} estimated as the ratio between the low and high frequency energy, the signal to noise ratio SNR of the current frame, the pitch stability counter pc representing the pitch period variation, the relative frame energy E_{rel} and the zero-crossing counter zc inside a frame. These parameters are considered together to build the merit function. The classification decision of the current frame is made depending on the value of f_m and also on the previous frame type.

The classification rules and also the calculation of all the parameters are thoroughly explained in [38]. Following the rules, the current frame is classified as VOICED, UNVOICED, ONSET and other TRANSITION frames. The silence frame is classified as UNVOICED here. The silence frame is identified by checking the VAD flag if operated in DTX mode, otherwise, the merit function is used. VOICED contains stable and periodic components. UNVOICED includes silence frame and is more like white noise. TRANSITION (including VOICED TRANSITION and UNVOICED TRANSITION) and ONSET are characterized by rapid variations of the energy. The speech classification on the word "success" is illustrated in Fig. 3.

B. Time-scaling in the excitation domain

The stretch and suppress operation for a speech sequence are illustrated in Fig. 4 (a) and Fig. 4 (b), respectively. The speech sequence can comprise VOICED, UNVOICED, TRANSITION and also ONSET frames, as illustrated in Fig. 3. According to the speech classification of each frame, VOICED and UNVOICED frames are processed differently. Moreover, some frames are not modified to prevent quality degradation, as proposed in [4]. Since pitch lag and pitch gain are internal parameters used by the *AMR* decoders, it is also advantageous to scale the speech inside the decoder, directly in the excitation domain. The different processing operations based on the result of speech classification are summarized as follows.

1) Time-scaling on VOICED: VOICED frames are extended or suppressed by adding or removing a certain number of pitch cycles to preserve the periodic nature. In the loop for each subframe in the analysis frame, information of pitch lags and pitch gains is first decoded, and also the adaptive codebook is updated before time-scaling to keep synchronization. The number of added or subtracted pitch cycles is determined by the difference between the original frame length and the expected frame length, combined with the pitch lag of the subframe. For extension, a certain number of pitch cycles are added just before the minimum energy point in the excitation signal, as shown in Fig. 5 (a). In the subframe with the maximum pitch gain, a search window of 20 samples is used to identify the minimum energy point P_{min}. For suppression, a certain number of pitch cycles are removed just before the minimum energy point P_{min} backwards, as shown in Fig. 5 (b). The minimum energy point P_{min} is found by searching in the last subframe. Fig. 6 (a) and Fig. 6 (b) show the extension and suppression of the same VOICED frame in the synthesis domain.

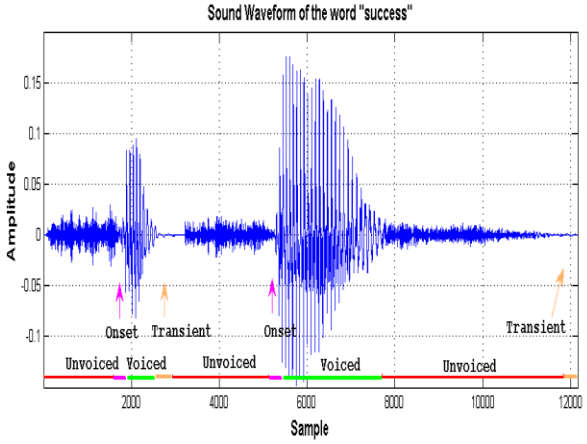


Figure 3. Speech classification of the word "success"

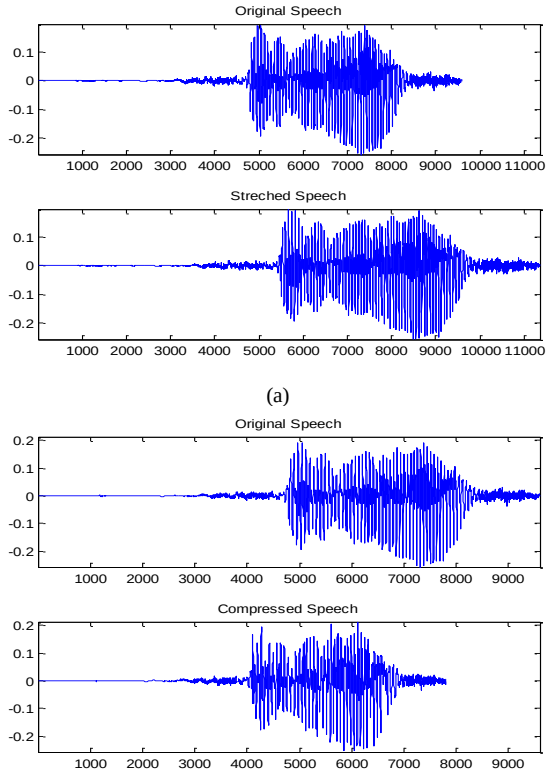


Figure 4. Time-scaling for speech sequence: (a) The speech sequence is streched. (b) The speech sequence is compressed

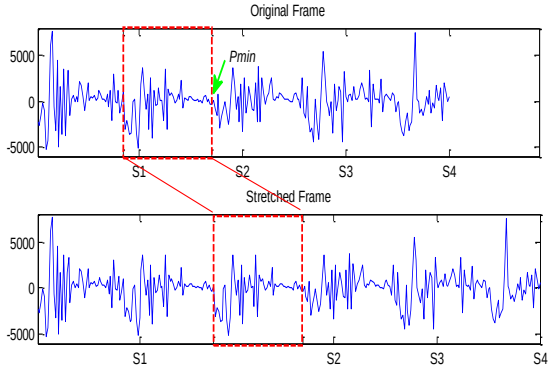
2) Time-scaling on UNVOICED: Time-scaling on UNVOICED frames is much simpler. For extension, a certain number of zeros are uniformly inserted in the excitation of the frame. In order to maintain the average energy per sample, a weighting factor which is the ratio between the requested and original frame lengths, is multiplied to the time-scaled excitation signal. For suppression, a certain number of samples are removed from the excitation of the frame. The samples can be removed from the beginning of the frame if the previous frame is UNVOICED or from the end of the frame if the previous frame is of other types. The number of zeros inserted or the number of samples removed relies on the expected new length, the original length and also the pitch lag.

3) Time-scaling on TRANSITION and ONSET: Since TRANSITION and ONSET frames contain components characterized by rapid variations, no time scale modification is operated on these frames to avoid artifacts.

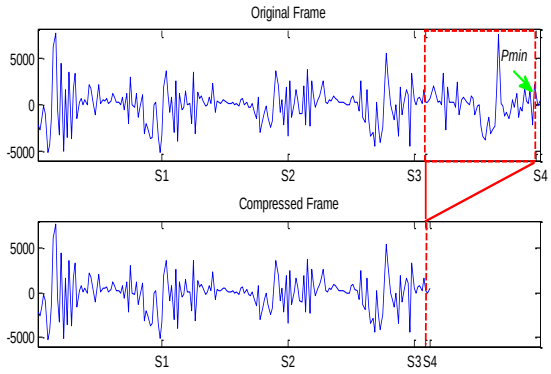
C. Modified CELP decoder

The modification of a simplified CELP synthesis decoder is illustrated in Fig. 7. Although the details can be different in each specific codec, the basic idea is quite similar. The generated excitation is formed by the fixed and adaptive codebooks with their corresponding gains. The excitation is classified into VOICED/UNVOICED/ONSET and other TRANSITION frames. According to the frame type decision, different time-scaling techniques are applied to the excitation signal. The reconstructed speech is obtained by feeding the scaled excitation of new length into the LP Synthesis Filter. In order to keep the synchronization between encoder and decoder, the adaptive codebook is updated before time-scaling.

In the *AMR-WB* decoder, the modification is a little bit more complicated, as the synthesis signal comprises two parts: the low-band synthesis SYN_{lo} and the high-band synthesis SYN_{hi}. In the loop of each subframe of the original length $L_{o,sub}$ (64 samples), the low-band excitation EXC_{lo} is formed by the fixed and adaptive codebooks with corresponding gains. Then the EXC_{lo} is post-processed and time scaled depending on the speech characteristics. The scaled low-band excitation is of the new length $L_{12.8,sub}$. The low-band synthesis SYN_{lo} is obtained by feeding the low-band excitation signal with the new excitation of length $L_{12.8,sub}$ to the synthesis filter and then performing up-sampling. Then the new low-band synthesis is of the new length $L_{16,sub} = L_{12.8,sub} \cdot 5/4$. The high-band synthesis is generated as usual but scaled with the new length $L_{16,sub}$. Finally, the high-band synthesis SYN_{hi} is added to the low-band synthesis SYN_{lo} to produce the synthesized speech signal SYN of the new length $L_{16,sub}$.

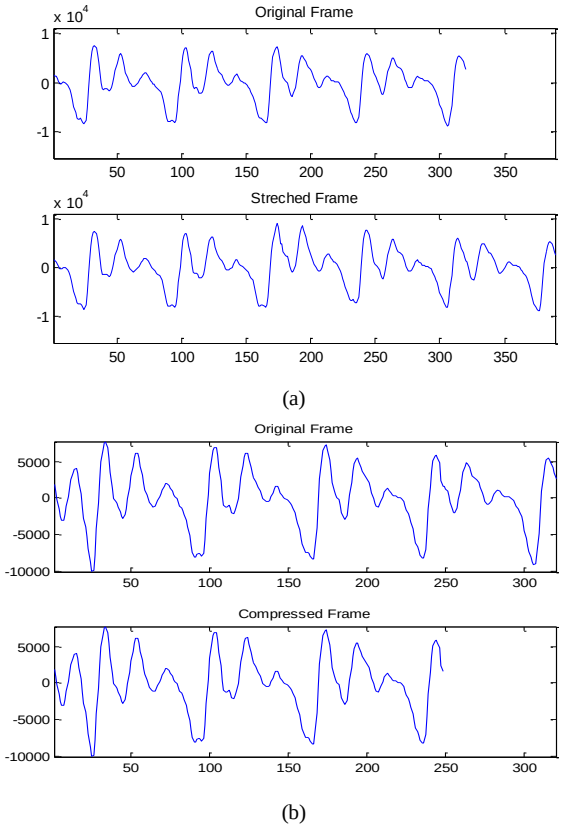


(a)



(b)

Figure 5. Time-scaling for VOICED frame in the excitation domain: (a) The frame is streched by adding one pitch cycle (b) The frame is compressed by removing one pitch cycle.



(a)

(b)

Figure 6. Time-scaling for VOICED frame in the synthesis domain: (a) The frame is streched by adding one pitch cycle (b) The frame is compressed by removing one pitch cycle.

V. Experimental Results

In the experiment, we implemented three other most promising algorithms, denoted as Algorithm 1 [2], Algorithm 2 [18], Algorithm 3 [4] to compare with our proposed jitter buffer algorithm. Our algorithm in this paper refers to Algorithm 4 for *AMR-NB* under random loss [1], Algorithm 5 for *AMR-NB* under bursty loss, Algorithm 6 for *AMR-WB* under random loss and Algorithm 7 for *AMR-WB* under bursty packet loss. The five traces used in the simulation are the same as used in [1]: each trace contains 7500 packets and the window size W used for both burstiness history and delay history is set to 300. The network statistics information for all five traces is shown in Table III. During the experiment, we implement the proposed playout delay estimation on the $I_{e,eff}$ model in (11) for Algorithm 5, (14) for Algorithm 6 and (15) for Algorithm 7. The maximum length after extension L_{max} is limited to twice of the original length (320 samples) and the minimum length after suppression L_{min} must not be shorter than half the original length (80 samples), as suggested in [4]. The maximum allowable end-to-end delay is 400 ms. The results for *AMR-NB* 12.2 kbit/s and *AMR-WB* 12.65 kbit/s are shown in Table IV.

From Table IV, it can be seen that Algorithm 5 achieves the highest R scores (consequently the highest MOS – see Section II) for *AMR-NB* and Algorithm 7 achieves the highest R scores for *AMR-WB* for all the tested traces. However, for narrow band Algorithm 4 has similar performance as Algorithm 5, and also for wide band the difference between Algorithm 6 and Algorithm 7 is minor. The performance comparison between Algorithm 4 and Algorithm 5 is shown in Fig. 8. We only show the results for trace 1 and trace 3 here, since they have the highest loss rate among all the trace files, as illustrated in Table III. The random loss model (Algorithm 4) and the bursty loss model (Algorithm 5) can lead to quite similar results when no packet or only a few packets are lost, i.e., $R \approx 1$. When the loss density is high (bursty), i.e., $BurstR > 1$, the difference can be perceived.

Algorithm 1 and Algorithm 2 are both talk-spurt based algorithms. Algorithm 1 estimates the playout time with the help of statistical delay information of several previous talk-spurts. Algorithm 2 implements an extended E-Model based on Weibull delay distribution. Both algorithms are not efficient for long talk-spurts and for cases where the network delay varies significantly such as in cases of spikes. Since the playout delay can only be updated in the next talk-spurt, the scheduled playout time cannot follow such spikes within a talk-spurt and results in more discarded packets due to late arrival, as for trace 1 and trace 5. Algorithm 3 and Algorithms 4 to 7 schedule playout delay on a per-packet basis. Algorithm 3 adjusts the playout time based on achieving an optimal trade-off between packet loss rate and end-to-end delay in a highly dynamic way and adapts more quickly to the network conditions even during speech activity (talk-spurt). But it does not provide a direct influence on the perceived speech quality, which is exactly the goal of the optimization. Algorithms 4 to 7 estimate the playout delay based on maximizing the MOS value derived from the rating factor R, therefore achieving the best performance in all the trace files.

The performance of playout delay estimation of trace 1 and trace 2 for *AMR-WB* is illustrated in Fig. 9. The results from Algorithm 3 and from our proposed Algorithm 6 are shown. Both algorithms adapt the playout delay quite well to the varying network delay. In the cases of spikes, our algorithm reduces the packet loss rate at the expense of additional delay.

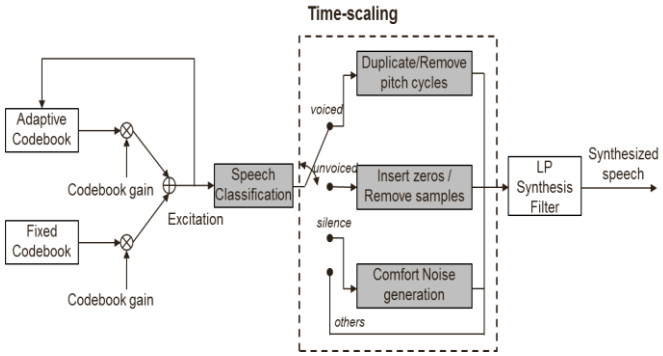


Figure 7. Modified architecture of the CELP synthesis model

TABLE III. TRACE FILE STATISTICS

Trace	Average network delay(ms)	STD of network delay(ms)	Average delay jitter(ms)	Maximum jitter(ms)	Network loss rate (%)
1	136.7	25.0	36.7	146	2.4
2	119.7	12.4	19.7	120	0.24
3	126.8	19.9	26.8	134	0.51
4	112.3	8.8	12.3	48	0
5	116.5	44.9	16.5	305	0

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed an adaptive jitter buffer algorithm based on the extended E-Model under bursty packet loss. We focused on the *AMR* codecs and a time-scaling technique embedded in the *AMR* decoders. Although the E-Model has limited accuracy in evaluating the speech quality, it can be used well in voice quality monitoring. Our simulation results show that the proposed method achieves better perceived speech quality compared to other existing algorithms under various network scenarios. Moreover, these results are not specific to *AMR-NB* and *AMR-WB*. As the time-scaling algorithm is closely related to the CELP coding scheme, the proposed jitter buffer management can be extended and adapted to other codecs, in particular to CELP based codecs.

For future activities, subjective listening tests are planned in order to validate the proposed method.

ACKNOWLEDGEMENT

We greatly appreciate the support and guidance of Dr. Herve Taddei, Dr. Christophe Beaugeant and Dr. Imre Varga. We would also like to thank Dr. Anisse Taleb and Mr. David Virette for their helpful comments and advice.

REFERENCES

[1] L. Pang and L. Böszörmenyi, "E-Model based Adaptive Jitter Buffer with Time-Scaling Embedded in AMR decoder," Proc. Int. Conf. on Digital Telecommunications, ICDT'11, April 2011, pp. 80-85.

[2] S. Moon, J. Kurose, and D. Towsley, "Packet audio play-out delay adjustment: Performance bounds and algorithms," ACM Multimedia Systems, vol. 6, no. 1, Jan. 1998, pp. 17-28.

[3] M. Narbutt and L. Murphy, "VoIP Playout Buffer Adjustment using Adaptive Estimation of Network Delays," Proc. 18th Int. Teletraffic Congress, ITC-18, 2003, pp. 1171-1180.

[4] Y. Liang, N. Färber, and B. Girod, "Adaptive Play-out Scheduling and Loss Concealment for Voice Communication Over IP Networks," IEEE Trans. on Multimedia, vol. 5, 2003.

[5] P. Gournay and K. Anderson, "Performance Analysis of a Decoder-Based Time Scaling Algorithm for Variable Jitter Buffering of Speech Over Packet Networks," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'06, May 2006.

[6] K. Fujimoto, S. Ata, and M. Murata, "Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications," Proc. IEEE Global Telecommunications Conf., GLOBECOM'02, vol. 3, Nov. 2002.

[7] L. Atzori and M. Lobina, "Speech playout buffering based on a simplified version of the ITU-T E-model," IEEE Signal Processing Letters, vol. 11, no. 3, March 2004, pp. 382-385.

[8] L. Atzori and M. Lobina, "Playout Buffering of Speech Packets Based on a Quality Maximization Approach," IEEE Trans. on Multimedia, vol. 8, 2006.

[9] ITU-T rec G.107, "E-model, a computational model for use in transmission planning," Dec. 2011.

[10] N.O. Johannesson, "The ETSI computation model: a tool for transmission planning of telephone networks," IEEE Communications Magazine, vol. 35, no. 1, Jan. 1997, pp. 70-79.

[11] A. Rix, J.G. Beerends, D. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality — Technology and applications," IEEE Trans. on Audio, Speech and Language Processing, vol. 14, 2006, pp. 1890-1901.

[12] A. Raja, R. Azad, C. Flanagan, and C. Ryan, "VoIP Speech Quality Estimation in a Mixed Context with Genetic Programming," Proc. 10th Annual Conf. on Genetic and evolutionary computation, July 2008, pp. 1627-1634.

[13] A. Raja, R. Azad, C. Flanagan, and C. Ryan, "A Methodology for Deriving VoIP Equipment Impairment Factors for a Mixed NB/WB Context," IEEE Trans. on Multimedia, vol. 10, no. 6, pp. 1046-1058, Oct. 2008.

[14] L. Carvalho, et al., "An E-Model Implementation for Speech Quality Evaluation in VoIP Systems," Proc. IEEE Symposium Computers and Communications, ISCC'05, 2005, pp. 933-938.

[15] S. Paulsen and T. Uhl, "Adjustments for QoS of VoIP in the E-Model," Telecommunications: The Infrastructure for the 21st Century, WTC, Sept. 13-14, 2010, pp. 1-6.

[16] W. Chen, P. Lin, and Y. Lin, "Real-Time VoIP Quality Measurement for Mobile Devices," IEEE Systems Journal, vol. 5, no. 4, Dec. 2011, pp. 538-544.

[17] K.S. Shanmugan, "Simulation-Based Estimate of QoS for Voice Traffic over WCDMA Radio Links," Proc. 5th Int. Conf. on Wireless Communications, Networking and Mobile Computing, WiCom'09, Sept. 24-26, 2009, pp. 1-4.

[18] L. Sun and E. Ifeachor, "Voice quality prediction models and their application in voip networks," IEEE Trans. on Multimedia, vol. 8, 2006.

[19] Z. Li, S. Zhao, and J. Kuang, "An improved speech playout buffering algorithm based on a new version of E-Model in VoIP," Proc. 3rd Int. Conf. on Communications and Networking in China, ChinaCom'08, 2008.

[20] L. Huang, Y. Chen, and T. Yaw, "Adaptive VoIP Service QoS Control based on Perceptual Speech Quality," Proc. 9th Int. Conf. on Advanced Communication Technology, 2007, pp. 885-890.

[21] C. Wu and W. Chang, "Perceptual Optimization of Playout Buffer in VoIP Applications," Proc. 1st Int. Conf. on

Communications and Networking in China, ChinaCom'06, 2006.

[22] C. Wu, K. Chen, Y. Chang, and C. Lei, "An Empirical Evaluation of VoIP Playout Buffer Dimensioning in Skype, Google Talk, and MSN Messenger," 18th Int. Workshop on Network and operating systems support for digital audio and video, 2009, pp. 97-102.

[23] L. Sun, G. Wade, B. Lines, and E. Ifeachor, "Impact of Packet Loss Location on Perceived Speech Quality," 2nd IP-Telephony Workshop, IPTEL'01, April 2001, pp. 114-122.

[24] L. Sun and E. Ifeachor, "New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks," Proc. IEEE Int. Conf. on Communications, ICC'04, vol. 3, June 20-24, 2004, pp. 1478-1483.

[25] R. Cole and J. Rosenbluth, "Voice over ip performance monitoring," ACM SIGCOMM Computer Communication Review, 2001.

[26] A. Raake, "Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions," IEEE Trans. on Audio, Speech, Language Processing, vol. 14, no. 6, Nov. 2006, pp. 1957-1968.

[27] H. Zhang, L. Xie, J. Byun, P. Flynn, and C. Shim, "Packet loss burstiness and enhancement to the E-Model," Proc. 6th ACIS Int. Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD'05, May 23-25, 2005, pp. 214-219.

[28] J. Liu and G. Wei, "Model the Packet-Loss Dependent Effective Equipment Impairment Factor in Speech Quality Estimation in VoIP and Its Realization," Proc. 2nd Int. Conf. on Advanced Computer Control, ICACC'11, vol. 4, March 27-29, 2011, pp. 359-362.

[29] 3GPP TS 26.071, AMR Speech Codec; General description, March 2011.

[30] 3GPP TS 26.171, AMR Wideband Speech Codec; General description, March 2011.

[31] B. Bessette, et al., "The adaptive multirate wideband speech codec (AMR-WB)," IEEE Trans. on Speech and Audio Processing, vol. 10, no. 8, Nov. 2002, pp. 620-636.

[32] ITU-T rec G.113, "Transmission impairments due to speech processing," Nov. 2007.

[33] ITU-T rec G.113 Amendment 1, "Revised Appendix IV-Provisional planning values for the wideband equipment impairment factor and the wideband packet loss robustness factor," March 2009.

[34] M. Goudarzi, L. Sun, and E. Ifeachor, "Modelling Speech Quality for NB and WB SILK Codec for VoIP Applications," Proc. 5th Int. Conf. on Next Generation Mobile Applications, Services and Technologies, NGMAST'11, Sept. 14-16, 2011, pp. 42-47.

[35] S. Möller, A. Raake, N. Kitawaki, and A. Takahashi, "Impairment factor framework for wideband speech codecs," IEEE Trans. on Audio, Speech, Language Processing, vol. 14, no. 6, Nov. 2006, pp. 1969-1976.

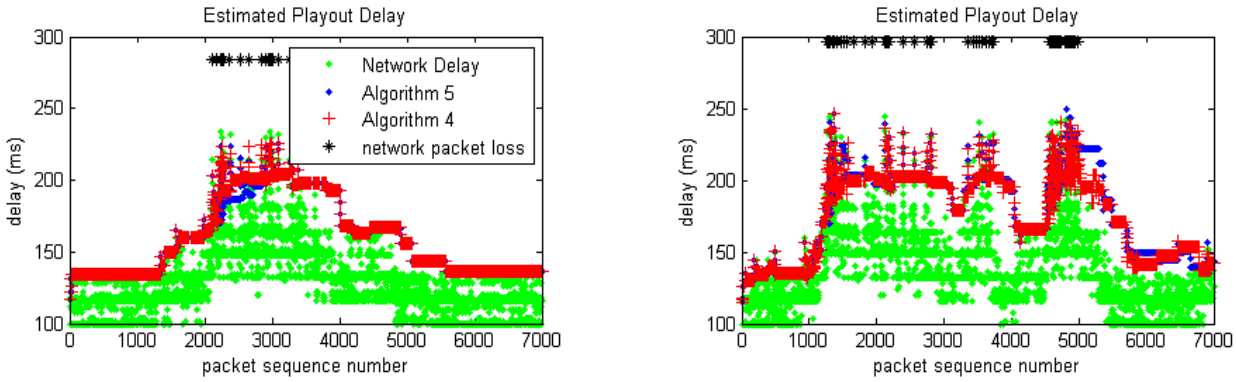
[36] S. Möller, N. Côté, V. Gautier-Turbin, N. Kitawaki, and A. Takahashi, "Instrumental Estimation of E-Model Parameters for Wideband Speech Codecs," EURASIP Journal on Audio, Speech and Music Processing, 2010.

[37] ITU-T rec G.109, "Definition of categories of speech transmission quality," Sept. 1999.

[38] 3GPP2 C.S0052-0, Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), June 2004.

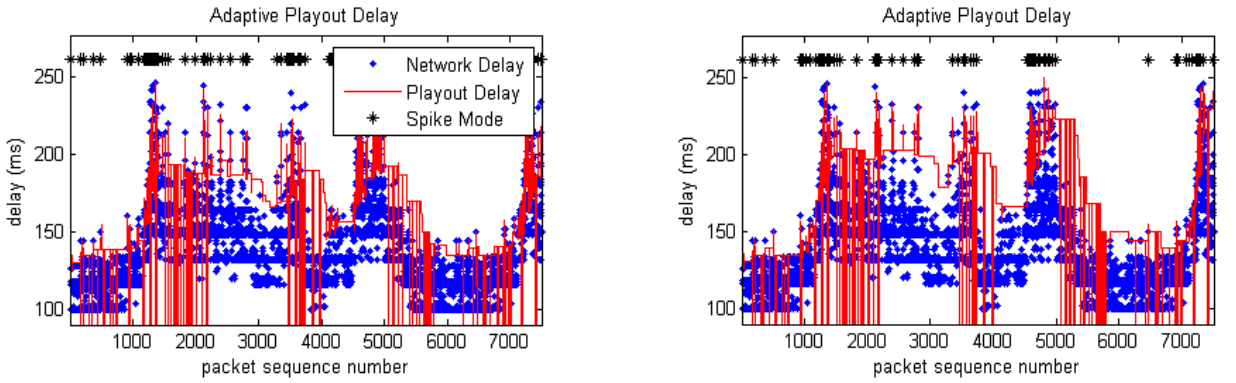
TABLE IV. COMPARISON OF DIFFERENT ALGORITHMS

	AMR-NB 12.2 kbit/s				AMR-WB 12.65 kbit/s			
Trace	Algorithm	Average playout delay(ms)	Late loss rate(%)	R	Algorithm	Average playout delay(ms)	Late loss rate(%)	R
1	Algorithm 1	180.5	7.5	39	Algorithm 1	180.5	7.5	40.5
	Algorithm 2	191.0	4.0	47.7	Algorithm 2	190.9	3.9	45.5
	Algorithm 3	165.2	3.5	51.3	Algorithm 3	165.2	3.5	48.1
	Algorithm 4	173.9	2.2	57.3	Algorithm 6	177.1	2.0	52.0
	Algorithm 5	175.9	1.9	57.3	Algorithm 7	176.8	2.1	52.1
2	Algorithm 1	153.3	2.3	66.1	Algorithm 1	153.3	2.3	59.9
	Algorithm 2	178.5	0.9	75.0	Algorithm 2	178.5	0.9	69.4
	Algorithm 3	150.4	1.5	71.0	Algorithm 3	150.4	1.5	64.7
	Algorithm 4	160.0	0.8	75.9	Algorithm 6	159.7	0.8	70.2
	Algorithm 5	160.0	0.8	75.9	Algorithm 7	159.7	0.8	70.2
3	Algorithm 1	148.6	6.0	49.2	Algorithm 1	148.6	5.0	46.8
	Algorithm 2	180.6	0.9	72.2	Algorithm 2	180.6	0.9	66.3
	Algorithm 3	154.7	1.2	71.5	Algorithm 3	154.7	1.2	65.2
	Algorithm 4	158.9	0.7	73.8	Algorithm 6	158.6	0.7	68.9
	Algorithm 5	157.9	0.8	73.8	Algorithm 7	158.6	0.7	68.9
4	Algorithm 1	133.7	0.3	82.3	Algorithm 1	133.7	0.3	78.2
	Algorithm 2	170.0	0.1	82.3	Algorithm 2	170.0	0.1	80
	Algorithm 3	134.7	0.4	81.7	Algorithm 3	134.7	0.4	77.3
	Algorithm 4	134.8	0.3	82.3	Algorithm 6	134.8	0.3	80
	Algorithm 5	134.8	0.3	82.3	Algorithm 7	134.8	0.3	80
5	Algorithm 1	147.6	2.6	66.2	Algorithm 1	147.6	2.6	60
	Algorithm 2	164.4	2.1	68.6	Algorithm 2	164.4	2.1	62.3
	Algorithm 3	146.0	1.2	75.1	Algorithm 3	146	1.2	69.1
	Algorithm 4	148.0	1.0	76.9	Algorithm 6	147.9	1.0	71.3
	Algorithm 5	147.9	1.0	76.9	Algorithm 7	147.9	0.9	71.3

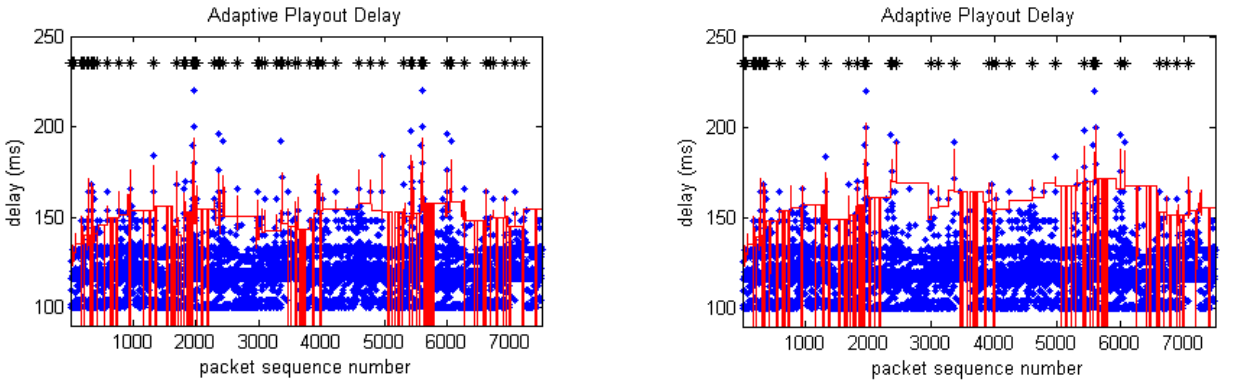


(a) (b)

Figure 8. Estimated Playout Delay based on Algorithm 4 and Algorithm 5 (a) trace3 (b) trace1



(a) (b)



(c) (d)

Figure 9. Playout delay estimation for Algortihm 3 and Algorithm 6: (a) Algorithm 3 for Trace 1 (b) Algorithm 6 for Trace 1 (c) Algorithm 3 for Trace 2

(d) Algorithm 6 for Trace 2

Enhancing Job Scheduling of an Atmospheric Intensive Data Application

Olivier Terzo, Lorenzo Mossucca, Klodiana Goga, Pietro Ruiu, Giuseppe Caragnano
Infrastructure and Systems for Advanced Computing (IS4AC)
Istituto Superiore Mario Boella (ISMB)
Turin, Italy
(terzo, mossucca, goga, ruiu, caragnano)@ismb.it

***Abstract*—Nowadays, e-Science applications involve great deal of data to have more accurate analysis. One of its application domains is the Radio Occultation which manages satellite data. Grid Processing Management is a physical infrastructure geographically distributed based on Grid Computing, that is implemented for the overall processing Radio Occultation analysis. After a brief description of algorithms adopted to characterize atmospheric profiles, the paper presents an improvement of job scheduling in order to decrease processing time and optimize resource utilization. Extension of grid computing capacity is implemented by virtual machines in existing physical Grid in order to satisfy temporary job requests. Also scheduling plays an important role in the infrastructure that is handled by a couple of schedulers which are developed to manage data automatically.**

***Keywords*-virtualization; grid computing; job scheduling; scalability; radio occultation; distributed environment.**

I. Introduction

Italian Space Agency (ASI) [1] developed a GPS receiver devoted to Radio Occultation (RO). The space-based GPS limb sounding, conventionally known as GPS Radio Occultation, is a remote sensing technique for the profiling of atmospheric parameters: refractivity, pressure, temperature, humidity and electron density, see [2], [3]. It is based on the inversion of GPS signals collected by an ad hoc receiver placed on-board a Low Earth Orbit (LEO) platform, when the transmitter rises or sets beyond the Earth's limb. The relative movement of both satellites allows a quasi vertical atmospheric scan of the signal trajectory and the profiles extracted are characterized by high vertical resolution and high accuracy. The RO technique is applied for meteorological purposes (data collected by one LEO receiver placed at 700 km altitude produce $300 \div 400$ profiles per day, worldwide distributed) since such observations can easily be assimilated into Numerical Weather Prediction models. Furthermore, it is also very useful for climatological purposes, for gravity wave observations and for Space Weather applications. This will cause the phase of the signal to be delayed. Moreover, the bent Geometric Optics trajectories followed by the signal during an entire occultation event will span the entire atmosphere in the vertical direction. As a consequence, through the inversion of the phase delay measurements, the refractivity related to each trajectory perigee can be evaluated, and a vertical profile can be identified. From refractivity, and adopting variational techniques, temperature and water vapor profiles can also be inferred. Given the characteristics of global coverage, good accuracy and high vertical resolution, products derived using such a technique are operationally used as input to weather forecasting model tools, and could also be harnessed in monitoring climate changes. ROSA-ROSSA (Radio Occultation Sounder for Atmosphere - Research Operational Satellite and Software Activities) is integrated in the operational ROSA Ground Segment it is operating at the ASI Space Geodesy Center, in Matera, Italy and at the Indian National Remote Sensing Agency [4], in Sriharikota, India. In this framework, Italian Space Agency has funded the development of the operational RO Ground Segment, which include the ROSA-ROSSA software. Partners of this project are several Italian universities, research centers and one industrial actor, which are responsible for the development and the integration of the various software modules defining the ROSA-ROSSA software: Istituto Superiore Mario Boella (Turin), Polytechnic of Turin, "Centro Interdipartimentale di Studi e Attività Spaziali" of Padua, University "La Sapienza" of Rome, University of Camerino, "International Center of Theoretical Physics" of Trieste, "Institute for Complex System" of Florence and "Consorzio per l'Informatica e la Telematica" of Matera.

The paper is structured as follows: Section 2 explains related work, motivation is shown in Section 3, Section 4 presents algorithms for Radio Occultation, Section 5 describes Grid architecture and virtual environment adopted to enhance capacity, Section 6 provides a brief overview about scheduling approach and its issue, Section 7 depicts some performance analysis and last Section draws conclusions and future work.

II. Related Work

The existent system is managed by an integrated software, called Grid Processing Management (GPM), devoted to handle and process data of the OCEANSAT-2 on board sensor. This architecture consists of the following components: physical worker nodes, repository, relational database, scheduler, agents and applications. The observed data, once acquired by the receiving ground station, are processed to produce refractivity, temperature and humidity profiles. The Radio Occultation events data processing consist of seven

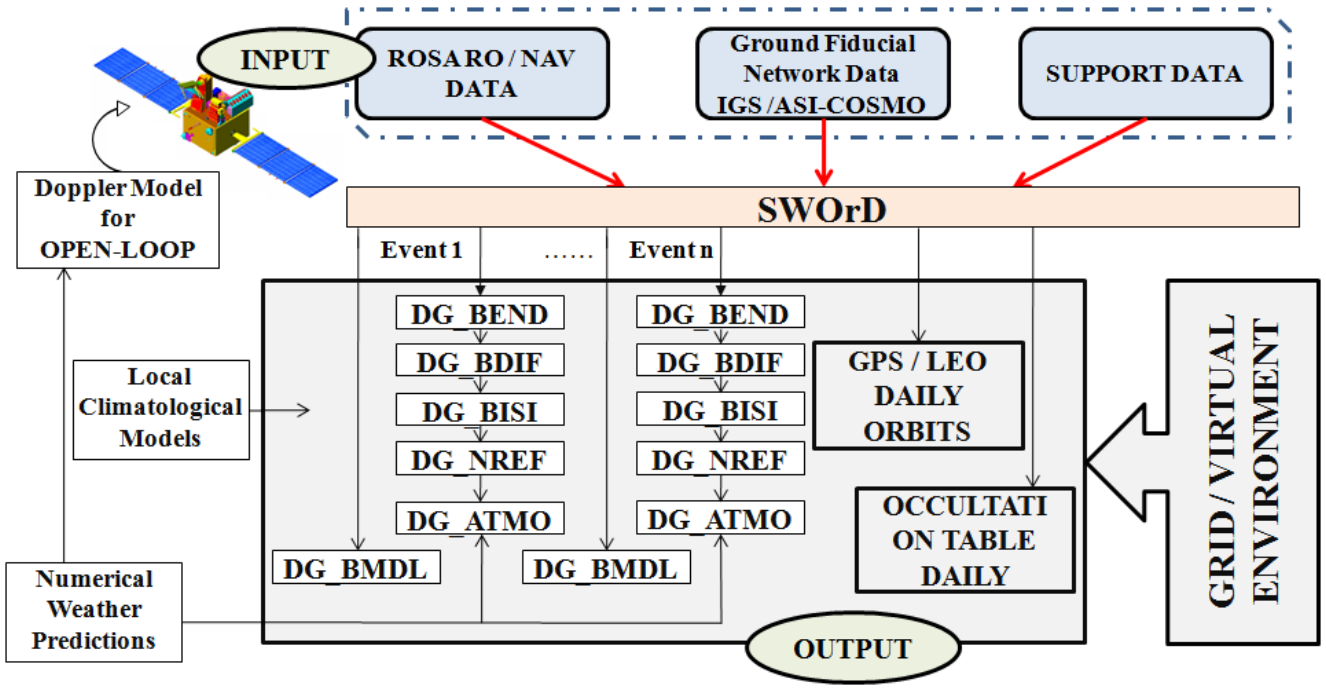


Figure 1. Data Generators Workflow.

main steps, named Data Generators (DGs). Figure 1 depicts the whole list of DGs that must be executed sequentially. Input and output data are daily composed of about 256 occultation events to be processed sequentially. For further details see [5]. In this context, where one needs to elaborate an enormous amount of data, using a grid architecture, there is already a great saving of time. In some cases the system fills up, when all worker nodes are elaborating data, increasing execution time. A solution to solve this problem can use a dynamically scalable system with virtual machines. The proposed architecture thus consists of a virtualized environment, which adds to the grid virtual nodes on demand, in order to increase the computational power and to solve load temporary peaks. In addition it allows to create multiple virtual nodes on the same machine optimizing physical resources, reducing energy consumption and decreasing maintenance costs. Virtualized systems help to improve infrastructure management, allowing the use of image template to create virtual nodes in a short time, speeding up the integration of new nodes on the grid and, therefore, improving reactivity and scalability of the infrastructure.

III. Motivation

The project aims to create a flexible architecture in order to manage the Radio Occultation data and to reduce their processing time. System guarantees the entire processing chain automatically that consists of seven DGs executed sequentially as explained before. During a learning phase, we evaluated that for each day, the events number to process are about 250, on a single machine the elaboration time for the entire chain processing, is approximatively 40 hours. The idea of using a distributed environment arose from the need to reduce this processing time because these make it difficult to get the results. The GPM has been developed with the goal of simplifying this task, by providing implementations of various core services provided by Globus Toolkit [6] and deemed essential for high performance distributed computing. Furthermore, it allows engineers and physicists involved to the project to have a common tool and infrastructure to process and share data, independently from the university in which they are. The chain process is composed of seven jobs developed by complex algorithms that involve a set of languages as Fortran, MatLab, C++, Mathematica, Java and Perl. Based on the limitations of software due to several programming languages involved in DGs algorithms, we decided to develop an ad-hoc job management scheduler stressing the importance of system scalability for grid infrastructure. A way to provide flexibility and scalability to the system is the implementation of a multi agents solution. The scheduler never gets information requiring the status to each nodes in fact on each node are installed two types of agents: Job and System agent. The first one is used to monitor the behavior of CPU, RAM and swap of nodes during the DG execution and it handles sending these information to master node. The second one, System agent, is used to monitor availability of each service on the node and periodically sends to the master node its own status: if all services

are available the node is ready to receive a job. The main advantage is that scheduler retrieves a pre list of available nodes ready for execution only with a simple query to the database.

IV. Radio Occultation Algorithms

ROSA-ROSSA software implements state of the art RO algorithms which are subdivided into seven different DGs executed in sequential mode [7], [8]. Starting from ROSA engineered data (or raw data observed by other RO payloads made available to the scientific community) coming from the ROSA on-board OCEANSAT-2 platform observations, from the ground GPS network (i.e., International GNSS Service network) and from other support data, the ROSA-ROSSA is able to produce data at higher levels, using a data processing chain defined by the following DGs:

- SWOrD;
- DG_BEND;
- DG_BDIF;
- DG_BISI;
- DG_NREF;
- DG_ATMO;
- DG_BMDL.

A. DGs Description

SWOrD is a software module that fully supports the orbit determination, orbit prediction, and which implements data generation activities connected with the ROSA sensor on-board OCEANSAT-2. Input data for SWOrD are ROSA GPS navigation and Radio Occultation observations, ground GPS network data and other support data. It generates the following output data:

- estimated rapid orbits and predicted orbits for the GPS constellation in Conventional Terrestrial Reference frame;
- estimated rapid orbits and predicted orbits for the OCEANSAT-2 platform in Conventional Terrestrial and Celestial Reference frame;
- 50 Hz closed-loop and 100 Hz Open-Loop excess phases and signal amplitudes data for each single occultation event;
- tables showing estimated and predicted (up to 6 hours in advance) occultation.

DG_BMDL predicts a bending angle and impact parameter profile usable as input in the ROSA on-board software Excess Doppler prediction module for open-loop tracking (see Figure 2). For each "predicted" occultation event, latitude and longitude of the geometrical tangent points (the nearest point of each trajectory to the Earth's surface, evaluated through predicted orbits) is used to compute bending angle and impact parameter profile from interpolated numerical weather prediction models (bending angle and impact parameter are geometrical parameter univocally identifying each trajectory followed by the RO signal). Inputs for DG_BMDL are predicted GPS and LEO orbits, respectively, and Predicted Occultation Tables, together with European Centre for Medium-Range Weather Forecasts (ECMWF) [9] world forecasts for the synoptic times valid for the future observed occultation event.

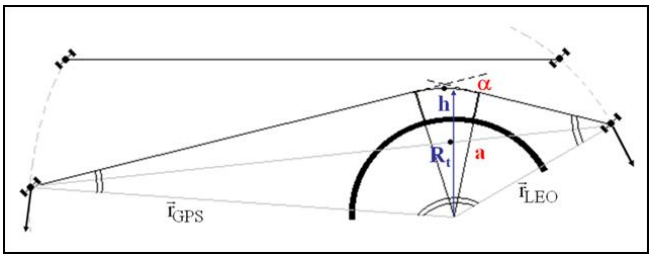


Figure 2. Radio Occultation Geometry.

DG_BEND provides raw bending angle and impact parameter profiles α(a) computed on GPS occulted signals on both GPS frequencies L_1 and L_2, by using a Wave Optics approach below a certain altitude. Above that altitude threshold, standard Geometrical Optics algorithms are applied. Inputs for DG_BEND are data L_1 and L_2 Excess Phases and signal amplitudes.

DG_BDIF provides (for each event) a bending angle and impact parameter profile, on which the ionospheric effects have been compensated for. This DG processes both L_1 and L_2 bending angle and impact parameters profiles given as input, in order to minimize the first order ionospheric dispersive effects. Outputs for DG_BDIF are bending angle and impact parameter iono-free profiles.

DG_BISI provides profiles of bending angle versus impact parameter optimized in the stratosphere above 40 km. In the ROSA-ROSSA, data coming from a Numerical Weather Prediction Model are used in place of climatological data for implementing the statistical optimization procedure necessary to reduce the high noise level left to the signal after ionospheric first order compensation previously applied by DG_BDIF.

DG_NREF provides (for each event) the refractivity profile and dry air temperature and pressure profiles. This DG is able to process iono-free and properly initialized bending angle and impact parameter profiles in order to compute the corresponding dry air "quasi" vertical atmospheric profiles.

DG_ATMO allows to evaluate the temperature and the water vapor profiles using forecasts or analysis obtained by numerical weather prediction. This DG receives on input from DG_NREF data files and produces on output data files, which contain the total temperature and total pressure

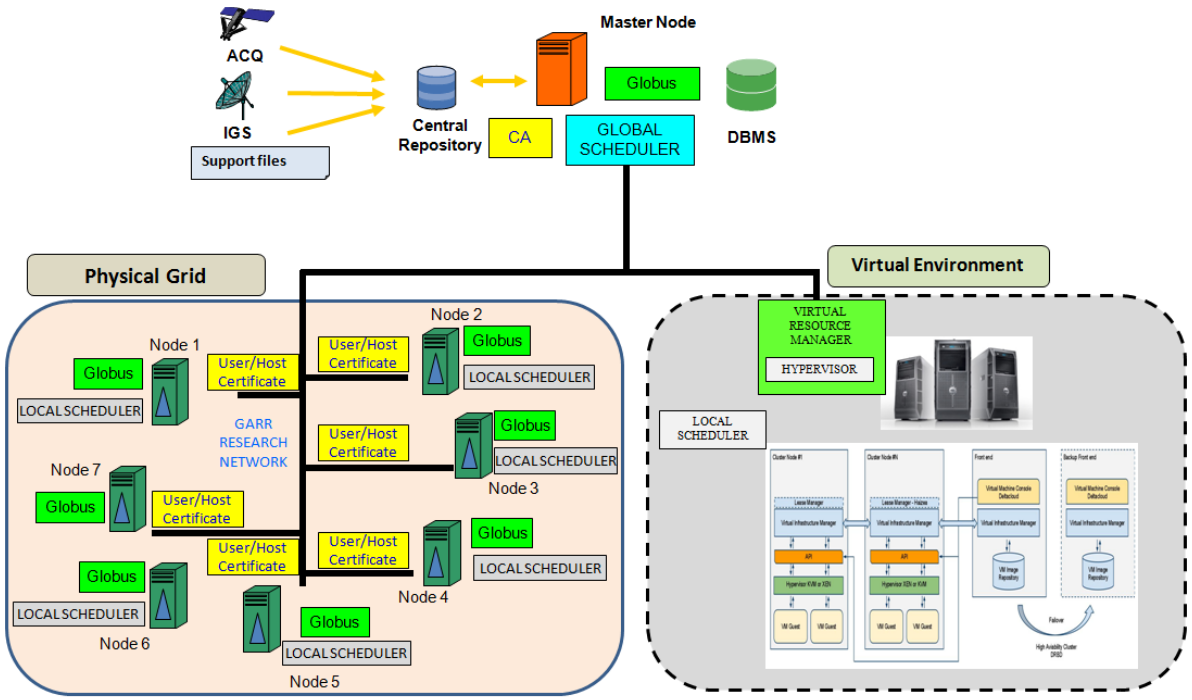


Figure 3. System Architecture.

profiles in terms of wet and dry components.

V. Architecture Design

The Grid Processing Management is an integrated system devoted to handle and process RO data of the OCEANSAT-2 ROSA on board sensor. The management software is developed in Java technology through the Java Commodity Grid (CoG) Kits. These allow to develop Grid application and administer Grids from a higher-level framework. The kits allow for easy and rapid application development and are used within the services of Globus Toolkit.

A. Grid Processing Management

The GPM is composed of the following subsystems (see Figure 3): middleware, central repository, relational database, scheduler, agents and DGs applications. The general purpose of our project is: sharing the computational resources, transferring a great amount of files and submitting jobs from several different organizations of the scientific community. Once input files stored, they are processed in an automatic way without any user interaction. The pool of resources consists of ten nodes: one Master Node and nine Worker Nodes, located geographically in Italy, each of them is equipped with a dual core processor, 2 GB of RAM and runs Linux x64_86 (Ubuntu distro). Master Node is responsible for the Data Base Management System, Certificate Authority, resource and job monitoring and job scheduling. Instead Worker Nodes contains grid and software tools used by jobs running which are received from Master Node.

The Globus Toolkit (GT) is a popular grid middleware [6], [10]. It provides a set of tools to create a grid infrastructure, covering security measures, resource location, resource management, communications which support the development of applications for high performance distributed computing environments, or computational grids [11], [12].

The main services included in GT are:

- Globus Resource Allocation Manager (GRAM): it converts a job request for resources into commands;
- Grid Security Infrastructure (GSI): for authentication of users and determines their access policies;
- Monitoring and Discovery Service (MDS): it collects information about resources such as processing capacity, bandwidth capacity, type of storage;
- Grid Resource Information Service (GRIS): it queries resources for their configuration, capabilities, and status;
- Grid Index Information Service (GIIS): coordinates arbitrary GRIS services;
- Grid File Transfer Protocol (GridFTP): provides a high-performance, secure, and robust data transfer mechanism.

The main reason for the choice of the GT is the availability under an open-source licensing agreement, which allow to use it freely and to improve the software. Grids need to support a wide variety of applications created according to different programming paradigms. Rather than providing a uniform programming model for grid applications, GT has

an object-oriented approach, providing a bag of services so that developers can choose the services that best meet their needs. These can also be introduced one at a time.

B. Extending Grid Capacity through Virtual Environment

Virtualization is a technology that allows running several concurrent operating system instances inside a single physical machine, reducing the hardware costs and improving the overall productivity by allowing users work on it simultaneously. The hypervisor, the fundamental component of a virtualized system, provides infrastructure support exploiting lower-level hardware resources in order to create multiple independent Virtual Machines (VM), isolated from each other. This virtualized layer, called also Virtual Machine Monitor (VMM), sits on top of the hardware and below the operating system. The hypervisor can control (create, shutdown, suspend) each VM that is running on top of the host machine. Multiple instances of different operating systems may share the virtualized hardware resources. The hypervisor is so named because it is conceptually one level higher than a supervisory program. A supervisory program or supervisor, also called kernel, is usually part of an operating system, that controls the execution of other routines and regulates work scheduling, input/output operations, error actions, and similar functions and regulates the flow of work in a data processing system (see Figure 4). Virtualization allows to gain significant benefits from the economic and the optimization of resources point of view [13]. Besides these, other noteworthy benefits are:

- security, stability and isolation: it is possible to run services in a virtual environment totally independent from each other;
- environmental impact reduction: optimization of resources implies reduction of power consumption and cooling;
- administration and management simplification: due to the common virtualization layer and the adoption of snapshots (installation and configuration);
- disaster recovery: VM can be started up in few minutes and can be cloned and distributed in different locations;
- high reliability and load balancing improvement: thanks to snapshots and live migration features.

Another aspect of virtualization is the adaptability: in fact it allows resource allocation to virtual hardware easily and quickly. We can maintain old servers with obsolete operating systems that cannot be moved to new servers as these OS would not be supported. In virtualized environments it is possible to run legacy systems allowing IT managers to get rid of old hardware no longer supported, and more prone to failure. In several cases it is appropriate to use virtualization to create test environments. It frequently happens that production systems need to be changed without knowledge about consequences, i.e., installing an operating system upgrade or a particular service pack is not a risk-free. Virtualization allows immediate replication of virtual machines in order to run all necessary tests.

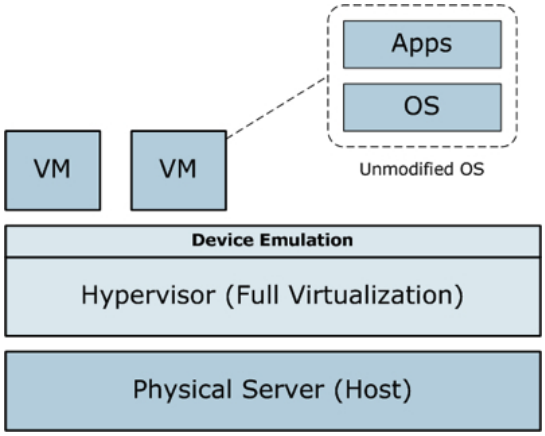


Figure 4. Full Virtualization.

C. Automatic chain

Our software allows to run the chain automatically and it is composed of two schedulers. The first runs on the Master Node, called Global Scheduler, checks for files ready for execution and sends them to Worker Nodes, according to well defined scheduling policies. The other one, called Local Scheduler, listens on the Worker Nodes, and when an input file is received, it is processed and output is returned to the Master Node [14], [20]. In Figure 5, data flow is depicted, the first transaction takes place on Master Node: it receives the files directly from the satellite and performs the first step of the chain, i.e., SWOrD, generating about 256 files that are placed in the input folder of the next step, DG_BEND. When there are files in the folder DG_BEND, the Global Scheduler checks available nodes by querying the database, and sends files to them. Global Scheduler takes care of automated scheduling of any input file. It uses all machines belonging to the Grid to distribute work load and to provide a backup system for all critical tasks within the system. The choice of how to share the file to run is based on 2 sets of scheduling policies, one concerning the available nodes and one derived from an analysis of the file to run. An Agent installed on each node, is used to monitor the availability of each service on the node. Periodically, it sends the general status of the node to the database on Master Node: if all services are active the node is in condition to receive a job. For the selection of Worker Nodes available and ready to run, the Global Scheduler checks on the database directly instead of querying each machine. When the Worker Node becomes aware of a file in its folder, the processing procedure starts. This will generate an output file that will be sent to Master Node in the next step folder, i.e., DG_BDIF. This procedure is performed for every steps of the chain, the operation is as

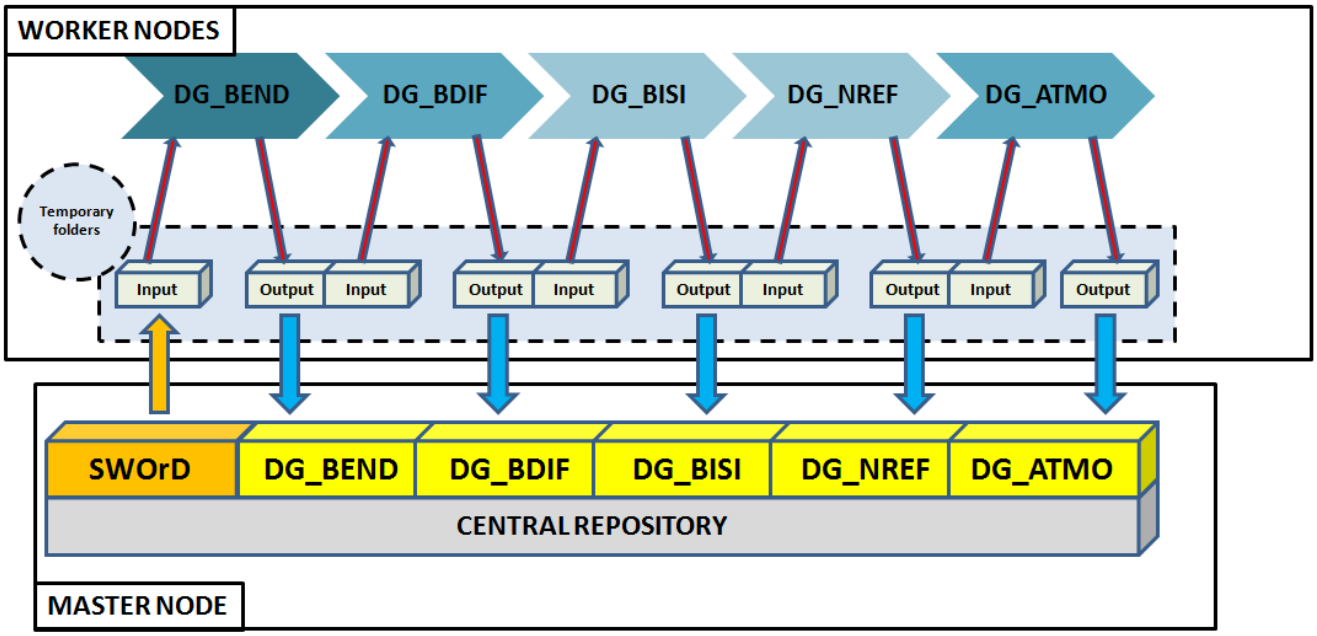


Figure 5. Chain Context.

follows: from SWOrD, the DG n-1 generates the output file that will be the input files of DG n, and so on. On Worker Nodes, each execution is performed in a temporary folder, so that, in case of error, it could be possible to identify the type of error made and then to reprocess the file. Two types of errors can occur: the first for lack of data in the file due to the satellite reception, the second for network failures or node crash. Only in the last case it is worth recover the process, and it is enough reprocess the input file. Furthermore, each process has a timeout, if within a fixed time processing has not been completed, the process is terminated in anyway. An important component of this architecture is the database, which allows to monitor any action of the Grid. Each transaction is stored on the database such as selected DG, start and end end time, input and output files, execution node and can contain type of error generated. The database also contains information on node status and if are available to receive the file to run, this allows to understand whether there are temporary network problems, so if the node is not able to receive the job.

VI. Scheduling Approach

The Grid scheduler selection is an important decision and significantly affects cluster utilization, availability, and intelligence. The most widely used are Portable Batch System, Torque scheduler, Maui Scheduler, Moab Workload Manager and Oracle Grid Engine, but they provide poor utilization of cluster's resources. These schedulers are enough closed and not are not easy to customize, in our system each machine is configured with different softwares and that means DG can not be executed on any Worker Node. This implies a most complicated job dispatching procedure that need of a high level of customization. The assignment of node for a new job execution is based not only on resource availability but also software needed for execution. To understand our choice to make a scheduler ad hoc in the following we want to give a brief overview of existing schedulers.

Portable Batch System (PBS)[15] is a networked subsystem for submitting, monitoring, and controlling a workload of batch jobs on one or more systems. Its main task is to allocate computational tasks, i.e., batch jobs, among the available computing resources. PBS is supported as a job scheduler mechanism by several meta schedulers including Moab and GRAM (Grid Resource Allocation Manager). With PBS it can specify the tasks to be executed; the system takes care of running these tasks and returns its results. If all computers are busy, then PBS holds your work and runs it when the resources are available. With PBS it is possible to create a batch job and then submit it. A batch job consists of a file containing the set of commands to run. It also contains directives which specify the attributes of the job, and resource requirements (e.g., number of processors and CPU time) that the job needs. Once PBS job is created, you can reuse it or modify it for subsequent runs.

Torque Resource Manager [16] provides control over batch jobs and distributed computing resources. Its name stands for Terascale Open-Source Resource and QUEue Manager. It is an open-source product based on the original PBS project and incorporates significant advances in the areas of scalability, reliability, fault tolerance, features extensions and functionality and is currently in use at tens of thousands

of leading government, academic, and commercial sites throughout the world. Torque can integrate with Moab Workload Manager to improve overall utilization, scheduling and administration on a cluster. The Torque Resource Manager is a distributed resource manager providing control over batch jobs and distributed compute nodes.

Maui Cluster Scheduler [17] is an open source job scheduler for clusters and supercomputers. It is a configurable tool capable of supporting an array of scheduling policies, dynamic priorities, extensive reservations, and fairshare capabilities. All of the capabilities found in Maui are also found in Moab, while Moab has added features including virtual private clusters, basic trigger support, graphical administration tools, and a Web-based user portal.

Moab [18] is a multi-dimensional policy-based workload management system that accelerates and automates the scheduling, managing, monitoring, and reporting of HPC (High Performance Computing) workloads on massive scale, multi-technology installations. The Moab accelerates both the decisions and orchestration of workload across the ideal combination of diverse resources, including specialized resources as GPGPUs. The speed and accuracy of the decisions and scheduling automation optimizes workload throughput and resource utilization so more work is accomplished in less time with existing resources to control costs and increase the value out of HPC investments.

Oracle Grid Engine [19], previously known as Sun Grid Engine (SGE), is an open source and free batch-queuing system, developed and supported by Sun Microsystems. SGE is typically used on a computer farm or high performance computing cluster and is responsible for accepting, scheduling, dispatching, and managing the remote and distributed execution of large numbers of standalone, parallel or interactive user jobs. It also manages and schedules the allocation of distributed resources such as processors, memory, disk space, and software licenses.

A. Job Management

The scheduling process takes place on the Master Node through the Global Scheduler (GS) [21]. This scheduler, developed in Java, is running on each folder dedicated to the DG, it waits for new events. Every time it takes notice of a new file to be executed, GS assigns and delivers it to the selected node according to predefined policies. The policies are split in the following steps:

- resources discovery is directly related to the information sent by the agents in order to return a pre list of available nodes;
- software discovery, by DG chose, selects the node that contains the software needed to run;
- status discovery devoted to check which nodes are free for execution (not running);
- hardware control: sorts the pre list based on computer power;

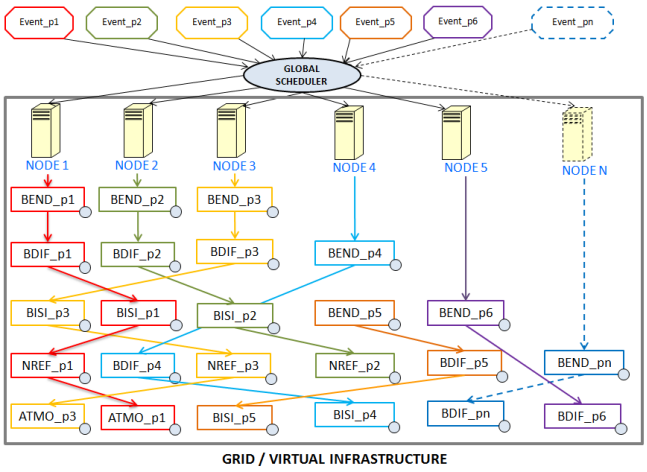


Figure 6. Dispatching Jobs.

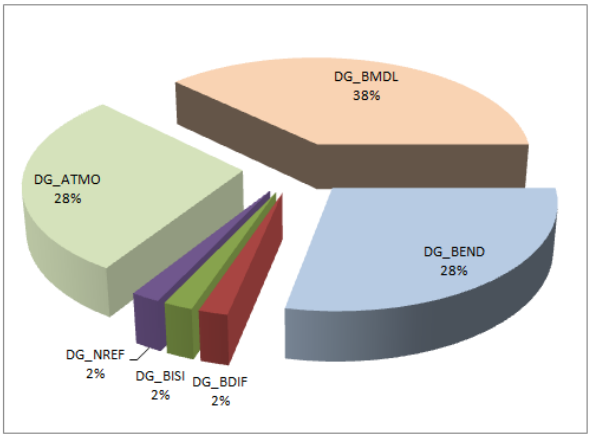


Figure 7. DG Percentage of Total Processing Time for a Single Event.

- queue control: at equality of power, it assigns job to node with less jobs in waiting.

VII. Improving Performances

Performance analysis have been executed for any DG, in order to improve scheduling policies. Since the data distribution is not homogeneous, this means that every hour SWOrD generates about 9 files which go in the folder DG_BEND and instead only 1 in folder DG_BMDL. Consequently for a day we have about 216 files for DG_BEND and 24 files for DG_BMDL. Figure 7 and Figure 8 represent the comparison for hourly and daily events, this allowed to assign a weight to each DG in order to calculate the amount of time when the CPU is busy.

The processing time of SWOrD step (about 72 minutes) is not included in these graphs because is executed only on Master Node therefore we can considered it outside from the grid. In order to compare performances between single machine and grid/virtual environment, we calculated processing time through the following Equations, varying

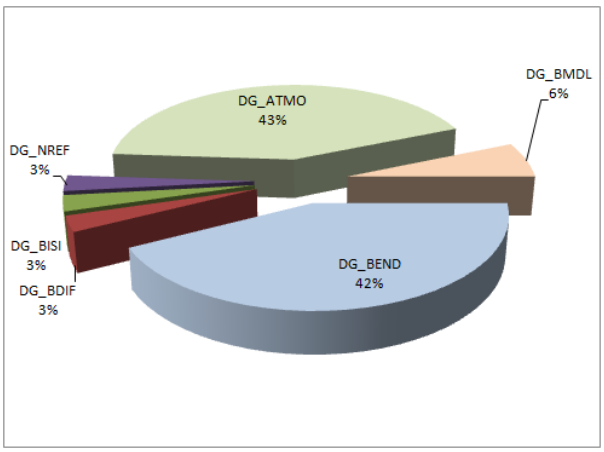


Figure 8. DG Percentage of Total Processing Time for Daily Events.

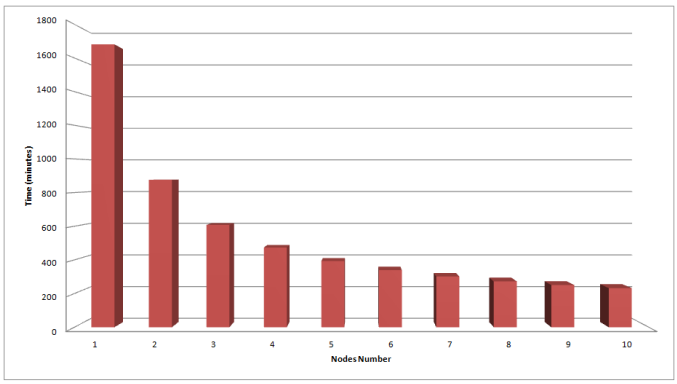


Figure 9. Estimated Processing Time for Daily Events (about 250 events)

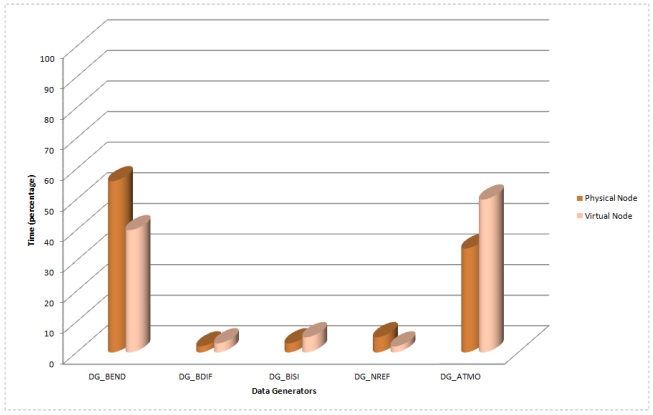


Figure 10. Comparison Physical Node vs Virtual Node.

nodes number. The Eq. 1 is referred to elaboration step by step sequentially on a single machine then $N = 1$.

$$T_{N=1} = T_e * \eta \quad (1)$$

While Eq.2 represents Grid/Virtual Environment that means $N > 1$. In this case, it is important to consider time for file transferring.

$$T_{N>1} = \frac{1}{N} \sum_{i=1}^{\eta} (T_{ei} + \beta) \quad (2)$$

Where: $T_{N=1}$ = Total Process Time on Single Machine, $T_{N>1}$ = Total Process Time on Grid/Virtual Environment, T_e = Event Process Time, T_{ei} = Event i Process Time, η = Number of RO Events, N = Grid Nodes Number, β = File Transfer Time.

Execution time trend is estimated in Figure 9, when the number of nodes and events is increased. When only a single machine is available, the total execution time for a set of daily files is 1752 minutes (about 29 hours), instead increasing the number of nodes, the execution decrease further, just note that with 2 nodes is 912 (about 15 hours). An important point when a single event is processed is that there is no gain time in grid environment respect to single machine execution; rather time is higher because we must consider the transfer time; it has a sizeable gain time only when there are a set of files to process.

Certainly, the benefits of the grid is ensure elaboration the overall chain in less time, instead, in distributed system where Worker Nodes are geographically located, it can have disadvantage in the network layer, in case of network failures or slow connections, to overcome this problem only internal nodes are always available for elaboration.

In Figure 10 we evaluated the elaboration time of each Data Generators executed on two types of nodes: physical and virtualized node. For DG_BDIF, DG_BISI, DG_NREF we can notice we have no elaboration time difference. Instead for DG_BEND and DG_ATMO we have a considerable time difference this can be due to different processor power (1.6 GHz for physical machines and 2.3 GHz for virtual machines). The server used is equipped with a dual-core Intel Xeon (4 CPU), 8 GB of RAM and 130 GB of storage. The virtual machines reside entirely on this server and therefore they share the resources (RAM, CPU, disk): each virtual node has 2 GB of RAM and 2 dedicated CPUs. They are configured exactly like a physical node of the grid with the same softwares and monitoring tools. It was decided to use Para Virtualized systems since it was shown that (in terms of network and I/O), they have better performances than the Full Virtualized one [22].

Para Virtualization Machine allows the operating system to be aware that it is running on a hypervisor instead of base hardware. The operating system must be modified to accommodate the unique situation of running on a hypervisor instead of basic hardware. The main advantage of this approach is the execution speed, always faster than HVM and Full Virtualization approach. The Xen hypervisor runs directly on the hardware and becomes the interface for all hardware requests such as CPU, I/O, and disk for the guest

operating systems. XEN is responsible for CPU scheduling and memory partitioning of the various virtual machines running on the hardware device. The hypervisor abstracts the hardware for the virtual machines and controls the execution of virtual machines as they share the common processing environment. It has no knowledge of networking, external storage devices, video, or any other common I/O functions found on a computing system.

VIII. Conclusion and Future Work

The ROSA-ROSSA software implements Radio Occultation technique, which was born for the first time on a Grid Computing infrastructure, called Grid Processing Management and in a second moment has been implemented a virtual environment to improve computing power and resources optimization. The project aims to be an example of application where users can use Grid Computing. In frameworks such as Radio Occultation, where the amount of data to be processed is significant, the use of a distributed architecture as the grid can be the optimal choice. We wanted to stress on a way to distribute jobs to nodes for execution in automatic way without any human interaction through a Local and a Global Scheduler. In order to increase computing capacity several virtual machines are added to the existing Grid infrastructure. As future work we are planning a further extension of the proposed architecture to clusters available across the European Grid Infrastructure (EGI) it provides access to high-throughput computing resources across Europe using grid computing techniques, furthermore we are studying a solution for Amazon Elastic Compute Cloud (Amazon EC2), that is a web service that provides resizable compute capacity in the cloud.

IX. Acknowledgment

The authors are grateful to the Italian Space Agency for supporting this project within the contract I/006/07/0 and to all ROSA-ROSSA partners for their contributions.

References

[1] Italian Space Agency(ASI), Available at: http://www.asi.it/, 2012

[2] Melbourne, W.G., Davis, E.S., Duncan, C.B., Hajj, G.A., Hardy, K.R. , Kursinski, E.R., Meehan, T.K., Young, L.E. and Yunck T.P.: *The application of spaceborne GPS to atmospheric limb sounding and global change monitoring*, JPL Publication, (1994), pp. 18-94

[3] Kursinski, E.R., Hajj, G.A., Schofield J.T., Linfield R.P., and Hardy K.R.: *Observing Earths atmosphere with radio occultation measurements using the Global Positioning System*, Journal of Geophysical Research 102(D19), (1997), pp. 23.429-23.465

[4] Indian Space Research Organization(ISRO), Available at: http://www.isro.org/, 2012

[5] Mossucca L., Terzo O., Molinaro M., Perona G., Cucca M., Notarpietro R.: *Preliminary results for atmospheric remote sensing data processing through Grid Computing*, The 2010 International Conference on High Performance Computing and Simulation, (2010), pp. 666-671

[6] The Globus Alliance, Available at: http://www.globus.org/, 2012

[7] Wickert J., Schmidt T., Beyerle G., Knig R., Reigber C. and Jakowski N.: *The radio occultation experiment aboard CHAMP: Operational data analysis and validation of vertical atmospheric profiles*, Journal of the Meteorological Society of Japan 82(1B), (2004), pp.381-395

[8] Luntama, J.P., Kirchengast, G., Borsche, M., Foelsche, U., Steiner, A., Healy, S., von Engeln, A., OClerigh, E. and Marquardt, C.: *Prospects of the EPS GRAS mission for operational atmospheric applications*, Bulletin of the American Meteorological Society 89(12): (2008), pp. 1863

[9] *European Centre for Medium-Range Weather Forecasts*, Available at: http://www.ecmwf.int/, 2012

[10] The Globus Consortium, Available at: http://www.globusconsortium.org/, 2012

[11] Foster, I. and Kesselman C.: *The Grid 2: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, San Francisco, CA, (2003) pp. 38-63

[12] Berman, F., Fox, G. and Hey A.: *Grid computing making the global infrastructure a reality*, Wiley, Chichester (2003), pp. 117-170

[13] *The Future Of Cloud Computing, Expert Group Report*, European Commmission, Information Society and Media, (2010), pp. 14-15

[14] Dimitriadou, S. and Karatza, H.: *Job scheduling in a distributed system using backfilling with inaccurate runtime computation, International Conference on Complex, Intelligent and Software Intensive System*, Washington DC, USA, (2010), pp. 329-336

[15] Portable Batch System, Available at: http://www.pbsworks.com/SupportDocuments.aspx, 2012

[16] Torque Resource Manager, Available at: http://www.adaptivecomputing.com/products/open-source/torque/, 2012

[17] Maui Cluster Scheduler, Available at: http://www.clusterresources.com/products/maui-cluster-scheduler.php/, 2012

[18] Moab, Available at: http://www.adaptivecomputing.com/products/hpc-products/moab-hpc-basic-edition/, 2012

[19] Oracle Grid Engine, Available at: http://www.oracle.com/us/products/tools/oracle-grid-engine-075549.html, 2012

[20] Xhafa, F., Pllan, S. and Barolli, L., *Grid and P2P Middleware for Scientific Computing Systems*, The international conference on complex, intelligent and software intensive system, 2010, pp. 409-414

[21] Leonid O., Rupak B., Hongzhang S. and Warren S.: *Job scheduling in a heterogeneous grid environment*, Lawrence Berkeley National Laboratory, (2004), Available at: http:// www.escholarship.org/uc/item/6659c4xj

[22] Kurowski, K., Nabrzyski, J.,A., Oleksiak, A. and Weglarz, J.: *Scheduling jobs on the grid multicriteria approach*, Computational Methods in Science and Technology 12(2), (2006), pp 123-138

Potentials and Limitations of CDMA Networks for Combined Inter-Satellite Communication and Relative Navigation

Rui Sun, Jian Guo, Eberhard Gill, Daan Maessen
Chair of Space Systems Engineering, Faculty of Aerospace Engineering
Delft University of Technology
Kluyverweg 1, 2629 HS, Delft, the Netherlands
r.sun@tudelft.nl, j.guo@tudelft.nl, E.K.A.Gill@tudelft.nl, D.C.Maessen@tudelft.nl

***Abstract*—Precision formation flying missions require formation acquisition and maintenance through the interactions among spacecraft by the inter-satellite communication and relative navigation. This paper analyses the dedicated system constraints of the network architecture for precision formation flying missions. The critical time issue and the operational flexibility are found to be two main constraints. Potentially applicable architectures are discussed and evaluated, which combine different multiple access technologies, half-duplex/full-duplex configurations and network topologies. It is proven that the most suitable and efficient architecture for PFF mission is the use of the half-duplex CDMA with the topology that allows the role of being reference rotating from one spacecraft to another in different time slots. The capability of CDMA is also investigated in terms of the multiple access interference. The paper verifies that this interference can limit the maximum number of spacecraft and bounds the inter-satellite range diversity. The interference exhibits a Doppler dependency and suffers as well from the near-far problem. Inter-satellite navigation accuracy will easily drop down below meter level at the moments of zero- or n-kHz Doppler crossovers, and also in case of the signals being corrupted by the near-by interferences. Two realistic mission scenarios are provided to verify the severe effects of the interference. Operational considerations and interference mitigation methods are also recommended.**

Keywords-precision formation flying; communication; relative navigation; CDMA; Multiple access interference

I. INTRODUCTION

Precision formation flying (PFF) missions involve the acquisition and maintenance of spacecraft in a desired relative geometric configuration, especially when trying to create a large virtual spaceborne instrument, such as the applications in the field of remote sensing and radio astronomy. Coordinating the components of such instruments on separate spacecraft can require highly accurate relative orientation and positioning [1].

Commonly, PFF missions make use of a differential Global Navigation Satellite System (GNSS) approach by exchanging GNSS-based navigation measurements via the inter-satellite links. Yet this method is limited to Low Earth Orbit (LEO). Many missions such as PROBA-3, Darwin, Magnetospheric Multiscale (MMS) and Terrestrial Planet Finder (TPF) [2-5] require the spacecraft flying in a High Earth Orbit (HEO) or Lagrange points, where GNSS signals are very weak or are even not available. As a result, a dedicated formation flying radio frequency (RF) technique using the locally generated inter-satellite ranging signals is necessary. The ranging capability is expected to integrate with the inter-satellite communication functionality for system efficiency. The use of the GNSS-like technology appears to be advantageous in such integrated system. Many missions that have been flown or proposed use this technology to assure a reliable inter-satellite communication and an accurate inter-satellite navigation, e.g., PROBA-3 [2], TPF [5], PRISMA [6] and MMS [7]. This paper also inherits the GNSS-like technology and regards it as a basic element to establish the PFF network.

In literature there are some discussions on the potential network architectures for formation flying. Bristow [8] proposed a concept called Operating Missions as a Note on the Internet (OMNI) that regards the spacecraft as network nodes and uses TCP/IP protocols to create a robust inter-satellite communications infrastructure. Similar proposals also include [9][10][11]. Vladimirova [12][13] discussed the potentials of applying WiFi or WiMax protocols for the establishment of the space wireless sensor network. An ability of implementing Ad-hoc has also been covered in order to support the high dynamics of spacecraft in large formations [14]. They all take advantages of the existing terrestrial protocols and try to move them to space. The benefits are the compatibility with the ground infrastructures and the good performance in terms of the large data throughput. However, the above two advantages are not the primary concern in PFF missions. The main requirement for the PFF inter-satellite link is to acquire and maintain the spacecraft in the desired relative geometry. The data exchanged should then be arrived timely to enable the estimations of the inter-satellite distance. This constraint is referred as to time critical issue in this paper and is elaborately analysed. Some of the existing terrestrial protocols will not be applicable due to the time critical issue. For example, WiFi, which functionally operates according to the detection of the medium [15], causes transmission uncertainty, which is a property that is unacceptable in PFF missions. This paper thus considers the medium access in fixed assignments of the possible connections by the classical multiple access (MA) technologies, such as Time, Code or Frequency Division Multiple Access (TDMA, CDMA or FDMA). Elaborate discussions on the choice of different MA technologies will be presented in this paper.

Another dedicated requirement for the PFF network is also proposed. It is the operational flexibility for

implementations across various mission phases. This requirement triggers the investigations on the choice of the network topologies and the half-duplex/full-duplex configurations. Considering all the above criteria, the paper will prove that the most robust and efficient manner in PFF missions is the use of the half-duplex CDMA with the topology that allows the role of being reference rotating from one spacecraft to another in different time slots.

The paper also discusses the limitations of implementing the CDMA concept. The well-known multiple access interference (MAI) will be introduced due to the non-perfect orthogonality of the Pseudo Random (PRN) codes. The Doppler dependency and the near-far problem are found to be two important factors that influence the MAI. The Dopplor dependency is severe when the Doppler offset between two spacecraft is close to zero or multiple integers of kHz. This is referred to as zero-Doppler crossover [16] and n-kHz Doppler crossover [17]. Server effects of the Doppler crossovers are investigated, especially in the situation where there is also a near-far problem. The resultant MAI error is analytically derived and verified using software simulations. Two case-study scenarios are used to determine the level of the MAI error and to predict the occurrence of the significant errors within an entire orbit period in realistic formations.

The paper shows several contributions, on the investigation of the possible network architectures, on the theoretical analysis of the selected CDMA strategy, on the analytical characterization of the MAI errors, and on the practical evaluation over two realistic missions.

The paper is organized as follows. In Section II, the inter-satellite communication and relative navigation system is introduced in order to propose the dedicated constraints for PFF network in Section III. Candidate network architectures considering different MA technologies, half-duplex/full duplex configurations, and network topologies are discussed and evaluated in Section IV. Then, network capability in terms of MAI is analysed, followed by an error characterization with respect to the Doppler offset and the near-far problem in Section V. In Section VI, two case-study scenarios are provided to identify the error level of MAI and to predict the occurrence of significant errors in realistic formations. Section VII concludes the paper and provides recommendations.

II. INTER-SATELLITE COMMUNICATION AND RELATIVE NAVIGATION SYSTEM

The inter-satellite system for PFF missions shall integrate communication and relative navigation into one package. Inter-satellite navigation is based on the locally generated inter-satellite ranging signals. A cost effective manner to generate these signals is to modify an existing GNSS receiver such that it can operate as a transceiver. The ranging method traces heritage to the GNSS-like technology and results in a highly miniaturized and accurate ranging device.

Such an inter-satellite system is expected to satisfy some specific high-level requirements according to the proposed missions [5][6]. According to these requirements and assuming the system operating in S-band 2.2 GHz, the EIRP (Effective Isotropic Radiated Power) is around 3 dB and the space loss is -129 dB. With the noise at the level of -134 dB of using 5 MHz bandwidth, it can be easily demonstrated that the received signals is 7 dB above the noise.

- Communication and relative navigation are integrated into one package, with mass less than 2 kg, and power consumption less than 2 W;
- Operating range less than 30 km;
- Omni-directional (4π steradian) coverage;
- Flexible to implement across various phases of mission operations in two modes: deployment, reconfiguration and collision avoidance tasks are achieved through the coarse-mode; formation maintenance in the desired configuration is required in fine-mode for scientific proposes;
- Navigation accuracy is in meter level for coarse-mode using code measurement (pseudorange) only, and centimetre level for fine-mode using the combination of code and carrier phases;
- Measurement update rate ≥ 1 Hz.

Space System Engineering group (SSE) in Delft University of Technology (TUDelft) has developed a software-defined inter-satellite communication and navigation simulator, which currently implements the signal generation and processing on PC via Matlab based on the system functional block in Figure 1. As shown, it comprises the transmitter and receiver front-end, signal generator and signal processing. The front-end comprises signal amplifier, band pass filter, down-conversion, sampling, and quantization. Frequency synthesizer is used in order to synchronize the edges of clocks in PRN code chipping frequency, and carrier and intermediate frequencies. The signal generator is used to generate carrier and GPS-like PRN code modulated signals. In single processing, acquisition, tracking and decoding modules are included for data recovery, code and carrier phase extraction, and pseudorange (-rate) derivation.

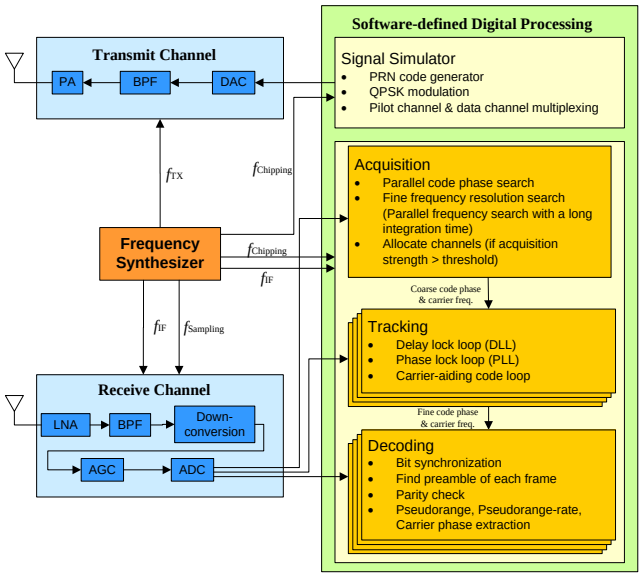


Figure 1. Inter-satellite communication and navigation system functional block diagram

Figure 2 depicts the signal generation and processing simulations assuming an intermediate frequency (IF) of 9.548 MHz and a sampling rate of 38.192 MHz. In the signal generator, the GPS C/A code is used as pseudorandom code,with a chipping rate of 1.023 Mcps and length of 1023 chips. The Doppler shift is assumed following a linear function relative to the simulation time.

The auto-correlation property, as shown in Figure 2(a), allows for accurate range measurements by yielding a sharp peak only when the replica code delay in the receiver is perfectly aligned to the code delay of the incoming signal. The signal energy is wideband spread due to the noise-like characteristic of the PRN code. PRN code spectrum is shown in Figure 2(b). The acquisition process, in Figure 2(c), is a global search in a two dimensional search space for approximate values of Doppler shift and code phase. This process is time and computation consuming. Therefore, parallel code/frequency one-dimensional search using Fast Fourier Transform (FFT) is implemented. After acquisition, control is handed over to the delay lock loop (DLL) and the phase lock loop (PLL). The fine estimates of the code and carrier phases will be obtained continuously after DLL and PLL. The variations due to the dynamics between spacecraft will also be tracked. As can be seen in Figure 2 (d) (c), in DLL, code error is ultimately reaching zero when the loop is getting to steady state, while in PLL, the linear function of Doppler is well represented. Communication bits in Figure 2 (e) are extracted from the tracking loops.

III. CONSTRAINS ON PFF NETWORK ARCHITECTURE

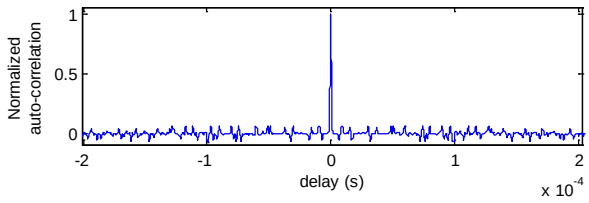
The time-critical issue and the operational flexibility are found to be two important criteria in PFF missions.

A. Time-critical requirement

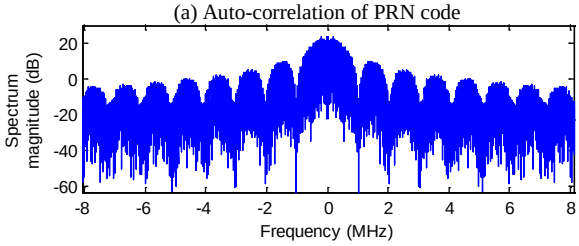
Time-critical requirements are driven by the nature of tight control, collision avoidance or scientific needs in some specific operational periods in PFF mission. These periods shall require high navigation accuracy and high measurement update frequency.

A relative navigation filter (e.g., extended Kalman filter) is used to account for the relative position errors resulting from all relevant non-modeled accelerations acting on the spacecraft. This process employs a numerical integration scheme in the filter that is updated at discrete intervals (t_i) as illustrated in Figure 3. The estimated relative state vector is obtained from an interpolation of the previous cycle. Based on all the measurements between t_i and t_{i+1}, a continuous polynomial representation of the trajectory is made available, which serves as starting point for the next filter update and relative orbit prediction [18]. Obviously the propagation period t_i-t_{i-1} has to be small for better approximations of the relative state vectors. On the other hand, t_i-t_{i-1} is limited by the processing time Δt_{proc}. Its typical value is 30 s for low earth orbits. In deep space, this period can be extended to several minutes or longer.

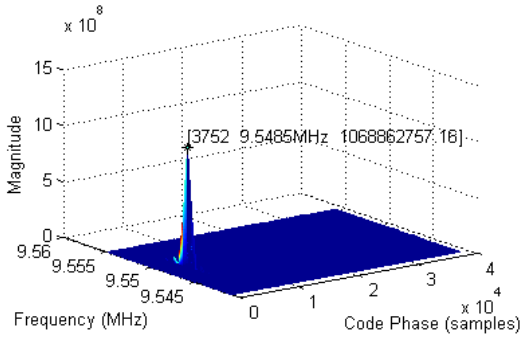
The measurements used in the filter are provided by the inter-satellite system. They can be either the unambiguous coarse code or the ambiguous precise carrier phases. Figure 4 exhibits their update timelines.



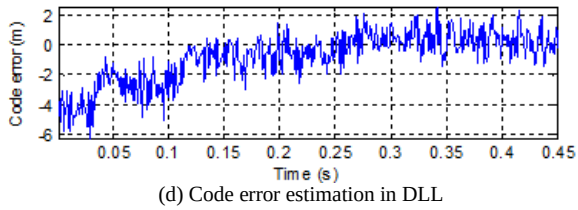
(a) Auto-correlation of PRN code



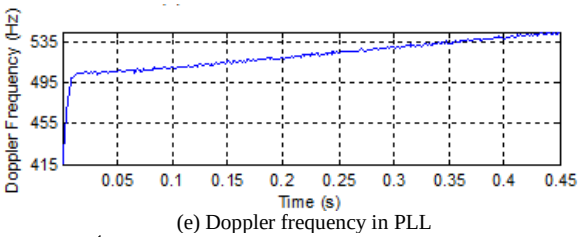
(b) PRN code spectrum



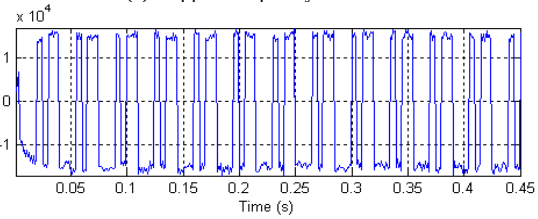
(c) Acquisition results. The peak is present at the point where the code delay is at 3752 sample (38192 samples in total for one complete PRN sequence) and carrier frequency is at 9.5485 MHz. That is, Doppler shift is approximately 500 Hz.



(d) Code error estimation in DLL



(e) Doppler frequency in PLL



(f) Communication bits extraction. The values are not ±1 because of the quantization in the front-end and integration in the tracking loops.

Figure 2. Software-defined signal generation and processing results

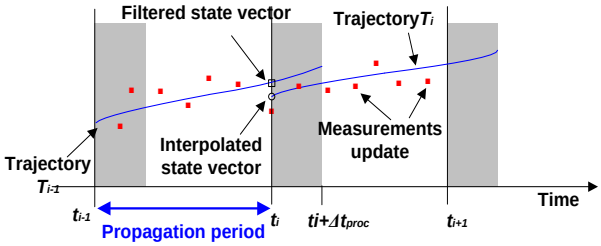
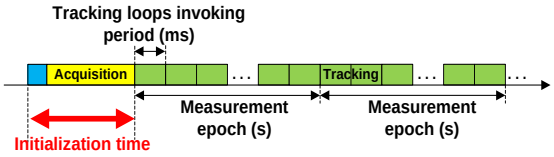
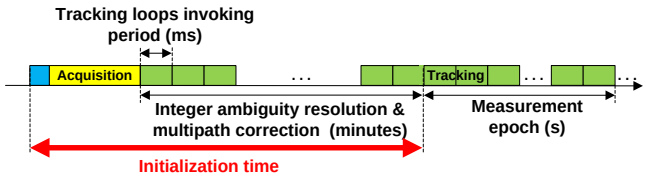


Figure 3. Timeline of relative navigation filter



(a) Timeline of code measurement generation



(b) Timeline of carrier phase measurement generation

Figure 4. Timeline of code and carrier phase measurements generation

It is observable in Figure 4 that the measurements are yielded after different initialization time, including the acquisition process for both of the codes and carrier phases, and some extra time only for the carrier phases to calculate the integer ambiguities and multipath corrections.

In the acquisition process, a long integration is performed to achieve sufficiently high carrier to noise ratio.

For integer ambiguities resolution, an sufficient large change of the relative geometry between spacecraft is normally required [6], which consumes long time. It is even more time consuming if the carrier phase itself is contaminated by multipath. For example, the integer ambiguity resolution used by PRISMA mission is to rotate a spacecraft for solving the line-of-sight (LOS) ambiguities firstly and then the distance ambiguities subsequently, taking 5 mins and 10 mins, respectively [6]. The resolution is combined with tabulated multipath correction through a filtering and smoothing process [6].

After initialization, tracking will be continuously running until the link is broken when switching the communication channel from one pair of spacecraft to another pair in a formation. At that time, re-initialization needs to be performed, including the corresponding initialization of sensor acquisition and carrier phase integer ambiguity resolution. This process consumes precious time and could result in a period that the on-board relative navigation filter in Figure 3 is propagating the dynamics for a couples of minutes without the measurement updates. The impact of such switching thus leads to an unsatisfactory performance of the navigation filter, which is unacceptable especially when the tight control is required at that specific time for the relative geometry maintenance.

This issue is referred to as time-critical constraint. The following PFF network architecture design should accommodate such time-critical constraint and give a high priority to timeliness instead of traditional considerations on the data throughput.

B. *Flexible operations across all mission phases*

Another important consideration of PFF networking is to recognize that the relative navigation requirements will change during the course of the mission's operations. The inter-satellite system is expected to operate across various phases of formation precision, requiring different levels of position sensing and control maneuvering.

Figure 5 illustrates the revolutionary phases of a PFF mission.

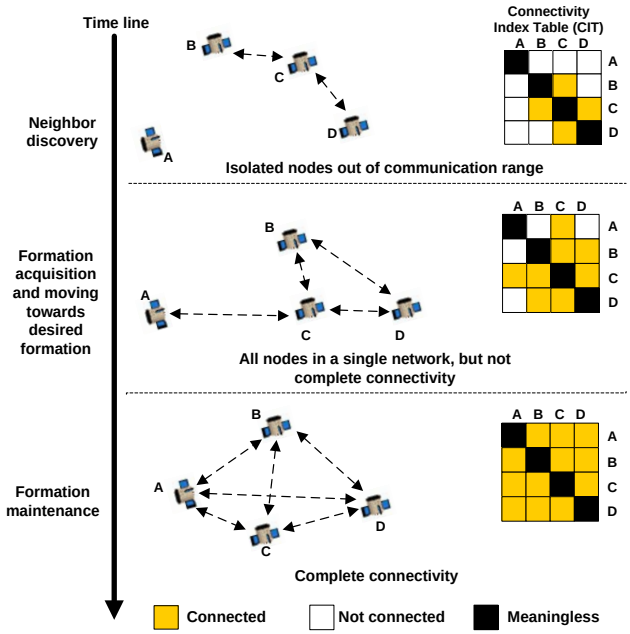


Figure 5. Evolutionary phases of a PFF mission

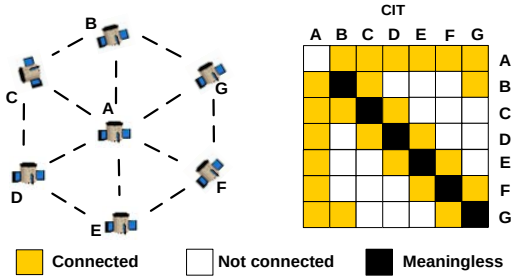


Figure 6. An example of complete connectivity with seven spacecraft in formation

In the initial deployment where the spacecraft are separated by substantial distances from one another, the resolution of position and orientation data are based e.g., on

coarse-mode sensors for collision avoidance, enabling further movement toward the desired configuration to take place safely. Spacecraft can be seen as free flyers located at a wide range of inter-satellite distances to each other. They will randomly access to the network. Neither a solely centralized nor a distributed topology is efficient in such situation, as some spacecraft are possibly out of the communication range of others.

As the spacecraft continue to aggregate into the desired spatial arrangement, they will eventually discover other spacecraft, which may be itself isolated or be already a member of a multi-spacecraft network. This condition, defined as formation acquisition and depicted in the centre of Figure 5, is in progress until all spacecraft are connected in a single network and moving towards desired formation.

Finally, when all spacecraft in the system show a "complete connectivity" and are settled into the desired pattern, formation maintenance are performed as shown in the bottom of Figure 5. Typically, a higher accuracy of their relative position is acquired, enabled by switching the inter-satellite system into the fine-mode. A precise formation may then be achieved using tighter control loops. Science operations will take place for e.g., multiple point remote sensing. At this moment, the mission topology will evolve to a centralized graph with one spacecraft at least chosen being the reference for a certain time period for relative navigation and formation control.

It is noted that in this final topology, "complete connectivity" does not mean all spacecraft must connect to each other, but connect to the desired spacecraft (normally the closed ones), which satisfies certain formation configuration. This happens especially in a large scale network such as depicted in Figure 6.

Subsequently, PFF can be reconfigured to set up a new arrangement for another science objective. The reconfiguration operation may drop back to coarse formation mode and prepares for new configuration, whereupon precise formation can again be executed [19].

A connectivity index table (CIT) is proposed to be part of traffic exchanged among spacecraft to share the current network condition. The measured range can also be included in the CIT, in which way a spacecraft will know all the relative positions even though not all of them are directly connected.

IV. NETWORK ARCHITECTURE FOR PFF

As a result of time-critical issue, networking solutions prefer the network nodes in a fixed assignment for all possible connections by MA technologies TDMA, CDMA, FDMA or their combinations, since they enable each spacecraft providing measurements from each of the others equally and timely [20].

In addition, the choices of half-duplex/full-duplex configurations and centralized/distributed topologies also play important roles in PFF network architecture.

Inevitably, if transmitting and receiving happen at the same time, some of the transmitted signal will leak into the receiver front ends and may easily saturate the receiver front ends or otherwise overwhelm the external signals. Half-duplex transceiver enables the transmitter and receiver taking turns to work, in which way "self-signals" are avoided. Full-duplex transceiver uses an appropriate filter to isolate the transmitter and receiver at their separated frequency bands to reject "self-signals". If the navigation measurements are not required simultaneously and continuously, half-duplex configuration is adequate and power-saving.

Network topology is expected to operate in a flexible arrangement, so as to account for the evolutionary phases of a PFF mission as shown in Figure 5. Neither solely centralized nor distributed topology is efficiently applicable during the neighbor discovery and formation acquisition phases, because some spacecraft are possibly out of communication range of others and could not access to the network. As the spacecraft progress towards the desired formation, it is better for the topology to evolve to a centralized graph in order to enable at least one spacecraft as reference for precise relative navigation and formation control. The role of reference can rotate from one spacecraft to another to avoid the problem of single point of failure.

Roles rotating at different time slots will give a robust and efficient connectivity. It can be implemented in a TDMA sequence with a strict timing boundary or a CDMA configuration with an adjustable period of time slot. Constraints of operating any of them come from the time-critical navigation requirements.

The possible combinations of different MA technologies, half-duplex/full-duplex configurations, and network topologies for PFF missions are illustrated in Figure 7-9.

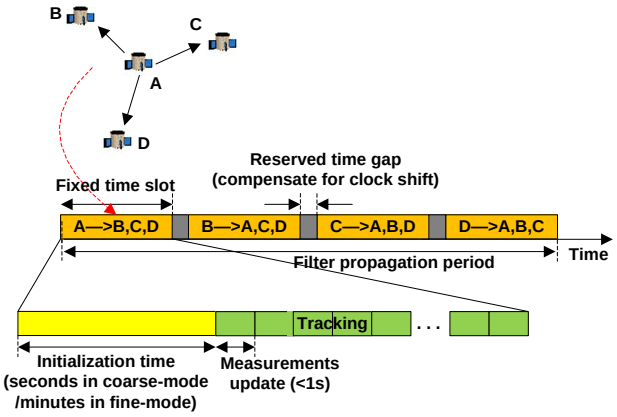


Figure 7. Half-duplex TDMA implemented for PFF mission

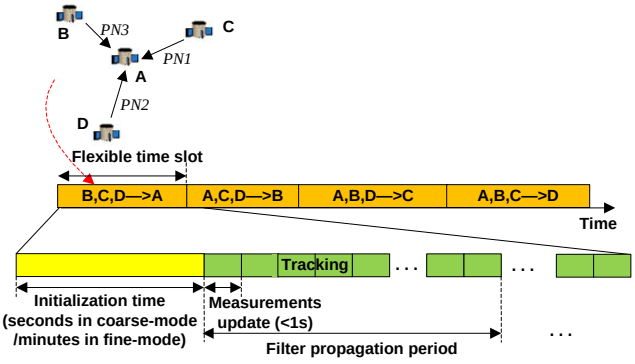


Figure 8. Half-duplex CDMA with roles rotation in PFF missions

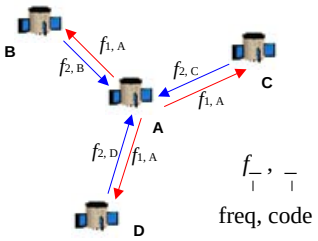


Figure 9. Full-duplex CDMA implented for PFF missions

In half-duplex TDMA mode in Figure 7, the measurements for propagating relative dynamics are from all of the spacecraft at different strictly fixed time slots. Therefore, the time slot is limited by the filter propagation period. That is, if a complete duty cycle of one TDMA sequence is as long as propagation period t_i-t_{i-1}, the time slot will be one-quarter of t_i-t_{i-1} (in four spacecraft formation), or even smaller to compensate for the clock drift by reserving a time gap between two slots. There is thus a high possibility that the time slot arranged to each spacecraft is too short to implement carrier phase measurement and the associated integer ambiguity resolution. This circumstance in TDMA mode can be improved by extending the propagation period, or equivalently, allowing the filter to freely propagate without measurement updates. However, the price of an increased relative navigation error has to be paid.

The fixed time slot of TDMA also limits the ability of working during different mission phases. It is caused by the variable inter-satellite range diversities and variable ranging accuracy requirements along with the course of the mission's operations.

Therefore, it is better to have an adjustable time slot and propagation period to avoid signal collisions and also guarantee efficient channel occupations and successful ranging abilities. Half-duplex CDMA with roles rotating architecture is thus proposed in Figure 8. It provides better capabilities in a way that the complete measurements from all of the spacecraft can be obtained in a single time slot by the use of CDMA strategy. This renders the duration of this time slot more flexible. It can be long enough to resolve the ambiguities and allow re-acquisitions. The implementations of either coarse-mode or fine-mode ranging at different mission phases are also possible as the time slot is adjustable. In addition, the signals transmitted from other spacecraft are not necessary to start at the same time in this CDMA mode, thus, it is tolerable if a spacecraft is joining in or dropping out of the formation.

Another advantage of such CDMA technology is that it can use the GNSS-like technology to the largest extent. For example, multiple channels in the system can work simultaneously. Differential measurements can be considered to improve ranging accuracy.

As comparison, another candidate architecture using full-duplex CDMA in centralized topology is shown in Figure 9. The isolation between transmitter and receiver is realized by separated frequencies and appropriate filters. It takes the advantages of allowing the signal transmissions among spacecraft without the necessity of re-acquisition each time at different time slots. Therefore, the measurements in this architecture can be processed simultaneously and continuously. The time-critical requirement is satisfied in an extreme solution that is the continuous connectivity. In addition, this method benefits from the fact that both clock offset and relative distance can be produced using dual one-way ranging [21].

However, the flexibility is low for this full-duplex CDMA strategy as it uses a centralized topology within the entire mission lifetime. The nature of full-duplex configuration needs more complicated system and consumes more power.

Considering all the above advantages and drawbacks for the three different architectures in Figure 7-9, half-duplex CDMA with roles rotating is the most robust and efficient manner for the PFF missions. The benefits of using it has been explained above. The following sections will focus on its limitations, including the well-known near-far problem and multiple access interference.

V. CDMA PFF Network Capability: Multiple Access Performance And Near-far Problem

The multiple access capability of CDMA is achieved by using the GNSS-like C/A code. However, it is not a completely orthogonal signalling format, which means cross-correlation is nonetheless present and induces noise in terms of MAI.

A. Cross correlation without Doppler effects

Assume that there are two signals, which are all un-correlated PRN codes with identical spectrum $G_s(f)$ and received at the same power level of P_s. The MAI term is introduced due to cross-correlation (CC) between undesired signal $c_m(t)$ and desired reference signal $c_k(t)$, where $c(t)$ represents PRN code. Ignoring the data modulation, Doppler frequency differences and noise for the moment, the MAI term is $c_k(t-\tau_k)c_m(t-\tau_m)$ with code delay τ_k, τ_m. Its power spectrum $G_{MAI}(f)$ is thus obtained by convolving the individual signal spectrum $G_s(f)$ [22]:

$$G_{MAI}(f) = P_s \int G_s(f) G_s(v-f) dv \quad (1)$$

Only the MAI spectrum near $f=0$ is important because the correlation filters have a small bandwidth on the order of Hz. $G_s(f)$ is in the form of $sinc^2$, thus [22]:

$$G_{MAI}(0) = P_s \int G_s^2(v) v = P_s \int_0^{\infty} \left(\frac{\sin \pi f / f_c}{\pi f / f_c} \right)^4 df = \alpha \frac{P_s}{f_c} \quad (2)$$

where f_c is chipping rate, α is a coefficient as a function of the filtered spectrum of $sinc^2$. If the spectrum includes all of its sidelobes, α is 2/3. If the spectrum is filtered to include only the mainlobe, α increases to approximately 0.815 [22].

Assuming *M* spacecraft at the same separation distances with the reference satellite in the formation, then *M-1* interfering multiple access signals exist. Considering white noise with noise spectrum density of N_0, the equivalent noise density and the effective energy per bit to equivalent noise density ratio are:

$$N_{0eq} = N_0 + \alpha(M-1)P_s / f_c \quad (3)$$

$$\frac{E_b}{N_{0eq}} = \frac{P_s T_d}{N_0 + \alpha(M-1)P_s / f_c} \quad (4)$$

where $T_d=1/f_d$, f_d is data bit rate. E_b/N_{0eq} determines the bit error rate. It is on the order of 10 dB if BER=10^{-5} and BPSK modulation without error correction coding is employed.

Furthermore, taking into account of the various separation distances between spacecraft, the near-far problem shows up. The effective E_b/N_{0eq} from a remote transmitter is further reduced due to the increase of MAI in close proximity. Because the free space loss is proportional to the square of distance, MAI spectrum density in eq. (2) is consequently multiplied by R_f^2/R_n^2, a factor to indicate the far desired signal to the near undesired interference range-squared ratio. E_b/N_{0eq} can then be revised to:

$$\begin{aligned}\frac{E_b}{N_{0eq}} &= \frac{P_s T_d}{N_0 + \alpha(M-1)(R_f^{\,2}/R_n^{\,2})P_s / f_c} \\ &= \frac{P_s}{N_0 f_d (1+\alpha(M-1)(R_f^{\,2}/R_n^{\,2})P_s/(f_c N_0))} \\ &= \frac{E_b}{N_0}\frac{1}{(1+\alpha(M-1)(R_f^{\,2}/R_n^{\,2})P_s/(f_c N_0))}\end{aligned} \quad (5)$$

The multiple access effect of *M-1* near interferences degrades the original E_b/N_0 by a factor of $1+\alpha(M-1)(R_f^{\,2}/R_n^{\,2})P_s/(f_c N_0)$. Note that signal to noise ratio P_s / N_0 equals to $f_d \cdot E_b / N_0$. Figure 10 displays the difference between the original E_b/N_0 and E_b/N_{0eq} caused by MAI effects. Assume using C/A code with chipping rate f_c of 1.023 Mcps, data bit rate f_d is 2 kbps, original E_b/N_0 is 10 dB, and coefficient α is 0.815 in case of 2 MHz bandwidth front-end filter by taking account of only the mainlobe spectrum.

As shown, in case of a small scale network and small distance diversity, e. g., only one or two interfering satellites in the near field of view, the difference between E_b/N_0 and E_b/N_{0eq} is less than 1 dB and can be negligible. However, as the number of satellites is getting larger or the far desired signal to the near interference range ratio is getting higher, this difference could be beyond the original E_b/N_0 threshold, which is unacceptable.

The impact factors of MAI, in the sense of the degradation of E_b/N_0, can be translated to:

- Maximum operating range diversity at a specific number of satellites in formation;
- Maximum achievable number of satellites in a formation that has low inter-satellite distance diversity.

For example, if five satellites in formation, one behaves like a reference and receives signals simultaneously from the other four. The maximum operating range is then in the sense that inter-satellite distance ratio could not be larger than 14 (red line in Figure10). On the other hand, if assuming the distance has only 3 times diversity, the maximum achievable number of satellites can be up to 620.

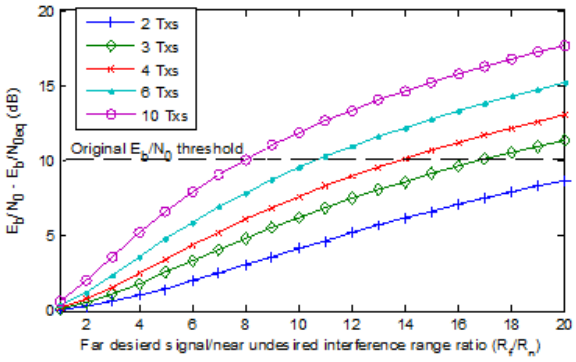


Figure 10. Difference between the original E_b/N_0 and E_b/N_{0eq} due to the near-far problem

Nevertheless, the results here are much more focused on communication performance and based on the assumption that Doppler effects on cross-correlations are ignored. When it comes to inter-satellite navigation, Doppler effect should be counted in, as significant errors will show up even in a small scale of network with low separation diversity.

B. Cross-correlation at high Doppler Offset

When taking into account of the different Doppler frequencies on the signals, there are Doppler frequency offsets on the cross correlations. The difference between two Doppler frequencies Δf_k, Δf_m for the desired and interference signals will be referred to as the Doppler offset $\Delta f_{m,k}$.

Without considering Doppler, the strongest and average cross-correlation peaks are approximately -24 dB and -30 dB lower than the main auto-correlation peak [23]. However, higher Doppler offset could exacerbate cross-correlation levels to -21.1 dB [23].

Specifically, for higher Doppler offset $\Delta f_{m,k}$, the cross-correlation term turns to

$$R_{CC} = \int_{(k-1)T}^{kT} c_k(t-\tau_k)c_m(t-\tau_m)\cos(2\pi\Delta f_{m,k}t + \Delta\phi_{m,k})dt \quad (6)$$

where T is integration time. The interference code structure $c_m(t)$ is changed by multiplication with $\cos(2\pi\Delta f_{m,k}t + \Delta\phi_{m,k})$. The effect of this multiplication can be significant when the Doppler offset is close to zero, then cross correlation errors become significant. This phenomenon has been observed and demonstrated by

researchers and was referred to as Doppler crossover [23][24].

Furthermore, PRN code (e.g., C/A) is not continuous but 1023 chip length sequence periodically repeated every 1 millisecond. That results in 1 kHz (f_c/P = *1.023Mcps/1023*) separated lines in the spectrum within *sinc*^{2} envelope [22]. Thus, for non-zero Doppler offset $\Delta f_{m,k}$, the spectrum of $c_m(t)$ is shifted with frequency $\Delta f_{m,k}$. If $\Delta f_{m,k}$ is not close to the integer multiple of 1 kHz, the line components of the desired signal spectrum and the shifted interference signal spectrum does not overlap, thus the cross correlation spectrum by convolving them $G_{m,k}(f) = \int G_k(f)G_m(v-(f+\Delta f_{m,k}))dv$ will diminish. On the contrary, mixing at the existing line frequencies on the order of several kHz will result in the interference being minimally suppressed. That is, if Doppler offset is an integer multiple of line component spacing 1 kHz, cross-correlation noise energy "leaks" through the correlation process, and could exacerbate cross-correlation levels to -21.1 dB [23]. The cross correlation errors in this case behave similarly to the zero Doppler crossover scenario, and have been regarded as n-kHz Doppler crossover [16].

In order to estimate the magnitude of the cross-correlation errors, the ranging system work process needs to be understood. As shown in Figure 1, in the tracking loops of DLL and PLL, three correlators early, prompt, and late I_E, I_P, I_L are used and feeding in the discriminator by calculating the difference between early and late correlators. Discriminator output then serves as feedback to adjust code delay and correlator outputs in a new round of iteration.

Discriminator output represents code tracking error. Under the assumption that auto-correlation is much larger than the cross-correlations, the normalized early-minus-late coherent discriminator can be written as [17]:

$$\begin{aligned} D &= (I_E - I_L)/(I_E + I_L)/2 \\ &\cong \left(R_{cc}(-\frac{d}{2}) - R_{cc}(\frac{d}{2}) \right) / R_{ac}(-\frac{d}{2}) / 4 \end{aligned} \quad (7)$$

where d is correlator spacing in chips, $R_{cc}(\pm d/2)$ and $R_{cc}(\pm d/2)$ are cross-correlation and auto-correlation with early or late delays.

It is obvious that the tracking error is related to the correlator spacing. Like the white noise, using smaller d, cross-correlation components $R_{cc}(d/2)$ and $R_{cc}(-d/2)$ become dependent and the common part could be canceled out.

Furthermore, combining eq. (6) and (7), the property of dependency on Doppler offset will show up. Simulation is given on the cross-correlation effect on the code tracking accuracy when Doppler offset $\Delta f_{m,k}$ linearly increased over time from -2500 Hz to 2500 Hz. The integration time used in the simulation is 20 ms, and correlator spacing is 0.8 chips. Only one interference signal is assumed, and has the same power level with the desired signal.

As displayed in Figure 11, zero and n-kHz Doppler crossover phenomenon are easily observed from the output of discriminator. However, different crossing points give different error contributions. At the zero or 2-kHz crossing moments, the reaction of code error is a sudden change of up to 5 m, while 1-kHz crossing point introduces error at approximately 2 m. The error pattern is a sinusoidal-like oscillation around the crossover point. A sensitive zone of ±25 Hz around the crossover point can then be defined where cross-correlation contributes the largest errors.

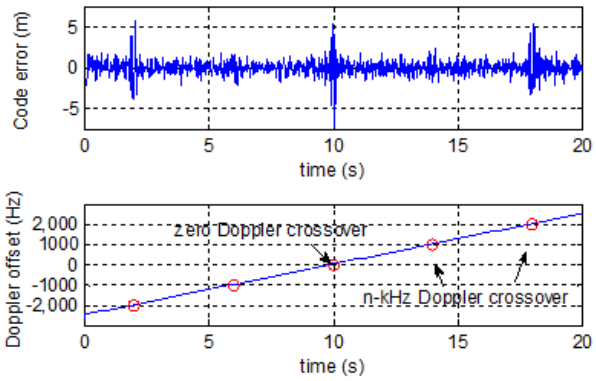


Figure 11. Code error estimation in presence of cross-correlation (top); Doppler offset between the desired and interfering signals, which is changing over time from -2500 Hz to 2500 Hz (bottom)

The cross-correlation error is not always as high as shown in Figure 11. According to eq. (6), it is also PRN code pattern and code delay dependent. RRN 7 and 22 are chosen with delay of 923 and 204 (1023 in total for C/A code) in this simulation after 1000 trials using Monte Carlo method. This code pattern and delay combination is at the moment when the average error plus standard variation is obtained.

It should be noted that the error magnitude is also affected by the code Doppler. Similar to the carrier Doppler offset, code Doppler offset also exists and will slowly change the relative delay between the desired and interfering signals, leading to the error magnitude being slightly enlarged or lessened. Randomly choosing code delays in Monte Carlo method can specifically illustrate the probability density distribution of the code Doppler effects in terms of code delay, which is however not shown in this paper in order to give a clear overview of the carrier Doppler crossover effect.

C. *Near-far problem at Doppler crossover*

The well-known near-far problem not only deteriorates the E_b/N_0 performance as displayed in Figure 11, but more seriously exacerbates the navigation accuracy, especially when it shows up at the moment of Doppler crossover.

Suppose the Doppler offset is 1 Hz, which is inside the sensitive zone of zero-Doppler crossover. Simulation in Figure 12 displays the errors when the interfering signal strength is linearly increased over time while the desired signal strength is constant. It is clear that cross-correlation error follows the sinusoidal-like oscillation at the frequency of 1 Hz, and the error magnitude also increases linearly.

That means the magnitude of cross-correlation error is proportional to the interfering/desired signal strength ratio. This relationship implies the same relevance between the

error and the inter-satellite distance diversity. Higher distance diversity in formation leads to poorer cross-correlations accordingly.

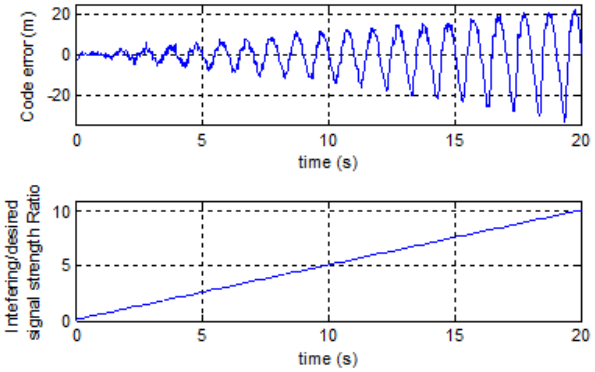


Figure 12. Code error estimation in presence of cross-correlation at 1 Hz Doppler offset (top); Inteference-to-desired signal strength ratio, which is changing over time from 0.1 to 10 (bottom)

VI. CASE STUDY: MAI EFFECTS IN FORMATION FLYING

Two case study scenarios are provided in order to evaluate the MAI effects in realistic formations.

A. *Low earth orbit chief-deputy formation scenario*

A formation geometry in the low earth orbit (LEO) with a chief satellite and several deputy satellites is a commonly used relative orbit geometry. In case of a Keplerian two-body motion, a circular chief orbit, and inter-satellite distance much smaller than the chief's semi-major axis, the relative dynamics can be expressed in Closhessy-Wiltshire (CW) equations in a linear form in Hill frame [25]. The relative motion of the deputy with respect to the chief is:

$$\boldsymbol{x} = (\boldsymbol{r}\ \boldsymbol{v})^T = (x\ y\ z\ \dot{x}\ \dot{y}\ \dot{z})^T \tag{8}$$

The vectors $\boldsymbol{r}$ and $\boldsymbol{v}$ denotes the relative positions and velocities in radial, along-track and cross-track directions. A safe ellipse relative orbit can be created and results in a closed form periodic solution when the initial orbit elements satisfy [25]

$$4x_0 + 2\dot{y}_0 / n = 0 \tag{9}$$

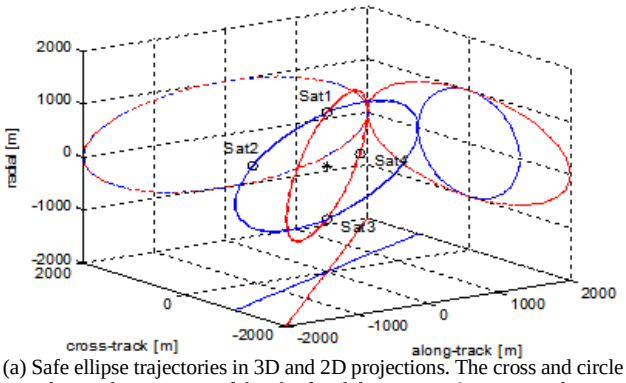
$$y_0 - 2\dot{x}_0 / n = 0 \tag{10}$$

where n is the orbital mean motion according to $n = (\mu/a^3)^{1/2}$, with μ the Earth's gravitational coefficient and a the semi-major axis of the chief. In such a safe ellipse orbit, the chance of collisions is minimized.

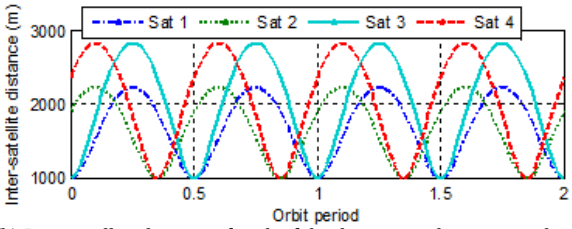
Suppose there are five satellites in formation, one is chief and the others are deputies in two safe elliptical orbits. The initial relative orbit elements are in Table I. The orbit of the chief is circular with a semi-major axis of 7000 km. After propagating the relative orbits using CW equations, it could be seen in Figure 13(a) that both of the relative orbits of deputy satellites are coplanar, which has an elliptical projections in the *xy*- and *xz*-plane, resulting in the linear motion in the *yz*-plane. The ellipse for Sat 1 and Sat 2 has the dimensions of 1×1×1 km, while Sat 3 and Sat 4 are in a 1×2×2 km ellipse. Inter-satellite distance of each of the deputies with respect to the chief is also displayed in Figure 13(b) in the pattern of sinusoid.

TABLE I. INITIAL RELATIVE ORIBT ELEMENTS

	x_0 [m]	y_0 [m]	z_0 [m]	$\dot{x}_0$ [m/s]	$\dot{y}_0$ [m/s]	$\dot{z}_0$ [m/s]
Sat 1	1000	0	0	0	-2.156	-0.178
Sat 2	587.78	-1618.03	-809.02	-0.872	-1.267	-0.634
Sat 3	-1000	0	0	0	2.156	2.156
Sat 4	-587.78	1618.03	1618.03	0.872	1.267	1.267



(a) Safe ellipse trajectories in 3D and 2D projections. The cross and circle denote the positions of the chief and deputies at t0, respectively.



(b) Inter-satellite distance of each of the deputies with respect to the chief

Figure 13. Safe ellipse relative orbit propagation in five satellite formation

When the chief satellite receiving signals from the deputies at the same time in CDMA architecture, multiple access interference will occur that results from their cross-correlation effects. As analysed in last section, cross-correlation is signal strength and Doppler dependent. Signal strength ratio between the interfering and desired signal could be easily calculated by inversely scaling the inter-satellite distance ratio. Doppler frequency is:

$$\Delta f_D = \frac{f_{carrier}}{c} v' = \frac{f_{carrier}}{c} \frac{\boldsymbol{v} \cdot \boldsymbol{r}}{|\boldsymbol{r}|} = \frac{f_{carrier}}{c} \frac{\dot{x}x + \dot{y}y + \dot{z}z}{\sqrt{x^2 + y^2 + z^2}} \tag{11}$$

Note that v' is not the satellite orbiting velocity, but the velocity projected in the inter-satellite link direction. Only

this part of velocity will introduce Doppler to the inter-satellite link. Carrier frequency of the signal $f_{carrier}$ is assumed to be 2.4 GHz in S-band, and c is speed of light.

For the receiver on the chief, multiple channels are allocated to track different PRN codes from deputies. Suppose one of these channels is for Sat 3, then signals from Sat 1, 2, 4 are regarded as interfering signals. Figure 14 provides their signal strength ratio and Doppler offsets, as well as the code errors resulting from the contributions of cross-correlations accordingly.

The period of Doppler offset or signal strength ratio is half of the orbit period. We only look at the first half. Note that during the whole period, Doppler offset is within or quite close to the Doppler crossover sensitive zone of ±25 Hz. That leads to a large error being brought in to the inter-satellite ranging system.

Analysing the bottom plot in Figure 14 it is shown that in general cross-correlation errors oscillate with changing frequencies and changing magnitudes. Maximum errors occur at the measurement batch of [0.3, 0.4] orbits and a clear slow frequency fading moment around 0.32 orbits is visible. This is caused by the exact zero-Doppler crossings of $\Delta f_{D,23}$ and $\Delta f_{D,43}$ that are happened coincidently at the moment of their highest signal strength ratio. Another zero-Doppler crossings for $\Delta f_{D,23}$ and $\Delta f_{D,43}$ occur around 0.02 orbits, but at a much smaller magnitude, because the signal strength ratio at that time is around nadir.

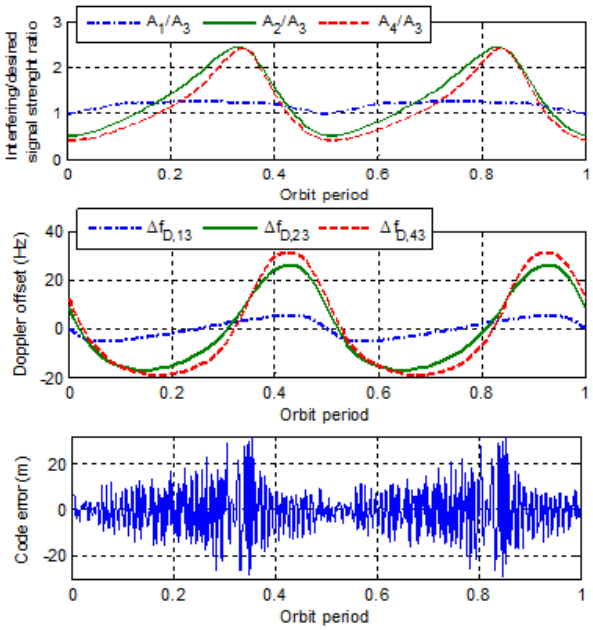


Figure 14. MAI effects in LEO safe ellipse formation geometry. Top: interfering/desired siganl magnitude ratio; Middle: Doppler offset; Bottom: code errors in presense of cross-correlations from 3 interfering Satellites.

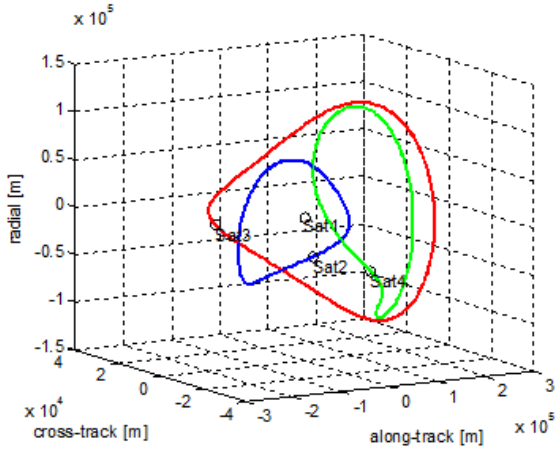
B. *Magnetospheric Multiscale formation scenario*

MMS formation is a NASA mission, which uses four identical spacecraft to make three-dimensional measurements of magnetospheric boundary regions and examine the process of magnetic reconnection [4].

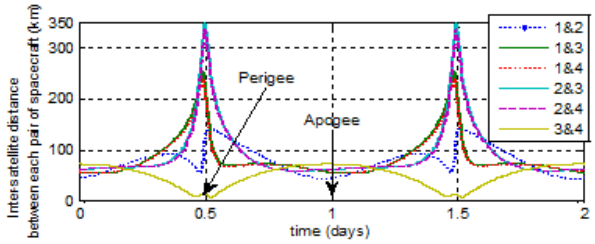
MMS mission has four identical spacecraft in a tetrahedral geometry. Inter-satellite communication will be conducted in a distributed topology and no signal spacecraft serves as chief. Two distinct phases will be in this mission. For Phase 1, MMS will be in a 1.2×12 Earth Radii highly elliptical orbit with a period of approximately one day. The initial orbital elements can be found in [4].

Unlike the low earth circular orbit in the last scenario, MMS mission cannot use linearized CW equations for relative orbit determination, but by propagating the absolute orbits of all the spacecraft using absolute Keplerian dynamics and then determining the relative motion in the Hill frame using a standard transformation.

Figure 15(a) displays the relative trajectories of Sat 2, 3, 4 with respect to Sat 1 to give a basic overview on how do they perform tetrahedral formation. Figure 15(b) exhibits their inter-satellite separation between each pair of spacecraft over two complete orbits. Near apogee, the inter-satellite distance are generally about 60 km, but near perigee, the spacecraft separations vary dramatically, which can be as large as 350 km and as small as 10 km.



(a) MMS mission relative trajectories in 3D of Sat 2, 3, 4 with respect to Sat 1. The circle denotes the positions of the four identical spacecraft at t_0.



(b) Inter-satellite distance between each pair of the spacecraft in MMS.

Figure 15. MMS mission relative orbit

MMS mission, with four purely identical and distributed spacecraft, is a good example of implementing CDMA with roles rotating architecture. At a flexible time slot, one of the spacecraft will be regarded as reference and receiving signals from the other three simultaneously. That indicates the potentials of multiple access interference. Furthermore, high distance diversity at perigee is particularly challenging for

CDMA architecture in the sense of near-far problem. It is valuable to investigate the most critical moments when the cross-correlation contributes the largest errors within a whole orbit period for MMS mission.

Figure 16 gives two examples of Sat 1 and Sat 4 as receivers, respectively, at two distinct time slots when interfering/desired signal arrived at receivers with dramatically different signal strength ratio.

The first example in Figure 16(a) shows maximum 4 times signal strength ratio, as well as the corresponding Doppler offset from minimum of -220 Hz to maximum of 800 Hz (blue dash line). Doppler offset is far beyond the Doppler crossover sensitive zone most of the time except for the crossing moments at around 0.20 and 0.92 days, when the slow frequency fading cross-correlation errors can be observed in the bottom figure. Interference at the duration of [0.38, 0.50] days, even with high signal strength ratio, should not contribute much more errors since the Doppler offset is extremely far from crossover. The visible errors at that duration actually come from the other interference contributor (green solid line) with its crossover at around 0.5 days and its general Doppler distribution all within the sensitive zone.

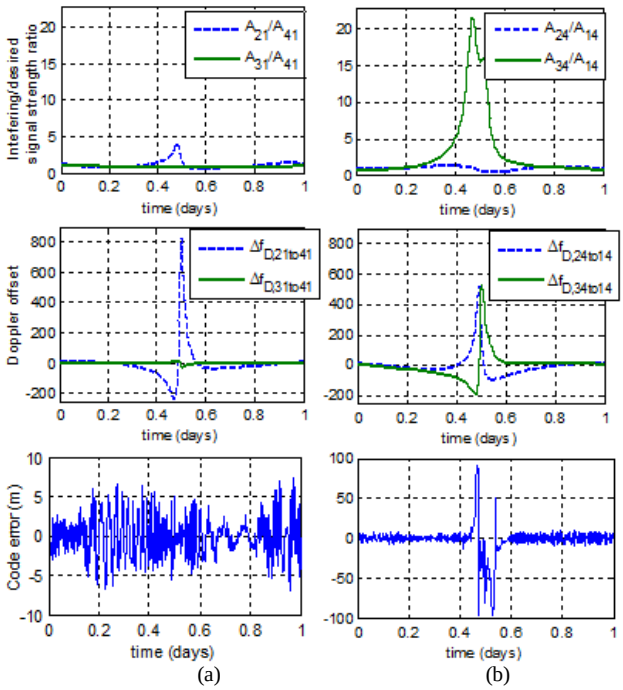


Figure 16. MAI effects in MMS mission. Top: interfering/desired siganl magnitude ratio; Middle: Doppler offset; Bottom: code errors in presense of cross-correlations when (a) Sat 1 is as receiver, requiring ranging signal from Sat 4 but interfered by Sat 2 and 3; (b) Sat 4 is as receiver, obtaining ranging signal from Sat 1 but interfered by Sat 2 and 3.

The other example in Figure 16(b) occurs more often because of the widely diverse inter-satellite separations, leading to severe near-far problem with up to 23 times signal strength ratio. This results in the degradation of the signal per bit to noise density reaching its bad performance threshold as explained before, but also yielding significant cross-correlation errors even if the corresponding Doppler offset beyond crossover sensitive zone. Up to 100 m errors are visible for a measurement batch of [0.45, 0.55] days. During that time, a Doppler crossover at around 0.50 days does happen, but luckily, is very instantaneous and almost does not leave even more severe consequences. The error magnitude is not proportionally increasing with the change of signal strength ratio, but keeping low due to the compensation by the large Doppler offset. The first error spike shows up when the signal strength ratio reaches 13. That indicates a kind of threshold for the occurrence of significant errors regardless of Doppler offset.

C. *Case-study summary*

The effects of near-far problem and Doppler crossover play important roles in defining the CDMA capability for the combined inter-satellite communication and navigation.

In case of low diverse inter-satellite separations, e.g., the scenario of safe ellipse orbit in LEO, the corresponding Doppler offset is diminutive, leading to high occurrence of zero-Doppler crossover and considerable MAI errors within the whole orbit period.

In case of high diverse inter-satellite separations, e.g., the scenario of MMS mission in highly elliptical orbit, Doppler offset is much higher and beyond the crossover sensitive zone for a substantial time of an orbit period, resulting in smaller cross-correlation errors. However, if the distance ratio between spacecraft reaches to the extent that even the large Doppler offset cannot compensate its influence any longer, significant errors will show up and affect the system accuracy to a large degree. To this end, an adaptive power control mechanism could be useful to minimize the impact of near-far problem, which ideally implements dynamically adjustable power attenuation in order to lower the transmitting power of interfering signals when they are in the close proximity.

Note that for both scenarios, there is no n-kHz Doppler crossover taking place. However, if thinking of going to higher carrier frequency, e.g., K-band or Ku-band, for the same orbit relative geometry, the chance of n-kHz Doppler crossover is getting high, which is also a resource of large MAI errors.

The methods of mitigating the MAI errors include an improvement of code delay loop inside the receiver by using a smaller correlator spacing or a longer integration time. One can also think of long time carrier phase smoothing of the pseudorange measurements, but should carefully take into account the requirements of high code and phase update rate in such PFF missions.

VII. CONCLUSIONS

In this paper, several network architectures are presented to support inter-satellite communication and relative navigation for PFF missions. Different multiple access technologies, half-duplex/full duplex configurations and network topologies are thoroughly analysed and combined as potential networking solutions for PFF missions. The

dedicated requirements for PFF missions are proposed, including the time critical issue and the operational flexibility, which are used as criteria to evaluate different architectures. Half-duplex CDMA with roles rotating is selected as a suitable architecture, as it enables system working with a wide range of flexibility, such as enabling both code and carrier phase measurements at variable mission phases, allowing to detect some spacecraft while tracking others, and being insensitive to a spacecraft joining in or dropping out of the formation. GNSS-like technology can also be utilized to the largest extent in CDMA network.

The limitation of using CDMA is investigated in terms of the Multiple access interference. This interference is Doppler dependent and suffers as well from the near-far problem. Regarding the communication performance, it is shown that the equivalent energy per bit to noise density ratio is reduced as compared to the case without MAI, leading to a limited inter-satellite separation diversity and a limited maximum number of spacecraft in the formation. Furthermore, MAI error worsens the navigation accuracy, especially at the moment of zero- or n-kHz Doppler crossovers or in case of signals being corrupted by the near-by interferences. Two case-study scenarios, one of a low earth circular orbit mission and another for a highly elliptical orbit mission, are provided that verifies the severe effects of MAI and the high probability of its occurrence within an entire orbit period.

MAI errors easily exceed the meter level, which can be mitigated using smaller correlator spacing or longer integration time. Long time carrier smoothing is also helpful to minimize the MAI errors. However, the requirements on the high update rates of the code and carrier phases need to be carefully taken into account in some tight control periods in the PFF missions.

REFERENCES

[1] R. Sun, J. Guo, E. Gill and D. Maessen, "Characterizing network architecture for inter-satellite communication and relative navigation in precise formation flying", 3rd International Conference on Advances in Satellite and Space Communications (SPACOMM 2011), Budapest, Hungary, 17-22 April, 2011.

[2] A. Wishart, F. Teston, et al. "The PROBA-3 Formation Flying technology Demonstration Mission", 58th International Astronautical Congress, Hyderabad, India, 24-28 September 2007.

[3] C. Bourga, C. Mehlen, et al. "A Formation Flying RF Subsystem for DARWIN and SMART-2", International Symposium Formation Flying: Missions & Technologies, Toulouse, France, 29-31 October 2002.

[4] M. Volle, T. Lee and A. Long, "Maneuver Recovery Analysis for the Magnetospheric Multiscale Mission", NASA Technical Reports Server, 2007.

[5] J. Y. Tien, J. M. Srinivasan, L. E. Young, et al, "Formation acquisition sensor for the Terrestrial Planet Finder (TPF) mission", IEEE Aerospace Conference, Big Sky, Montana, 6-13 March, 2004.

[6] V. Barrena, M. Suatoni, C. Flores, J. Thevenet and C. Mehlen, "Formation flying RF ranging subsystem for RPISMA: navigation algorithm design and implementation", Proc. 3rd Int. Symposium on Formation Flying, Missions and Technologies, Noordwijk, the Netherlands, 23-25 April 2008.

[7] G. Heckler, L. Winternitz, W. Bamford, "MMS-IRAS TRL-6 Testing", ION GNSS 2008, Savannah, GA, 16-19 Sep. 2008.

[8] J. Bristow, "A Formation Flying Technology Vision", AIAA Space Conference and Exposition, Long Beach, CA, 19-21 Sep. 2000.

[9] K. Hogie, E. Criscuolo and R. Parise, "Using standard Internet Protocols and Applications in Space", Journal of Computer Networks, vol. 47, 2005, pp. 603-650.

[10] L. Wood, W. Ivancic, D. Hodgson, et al, "Using Internet Nodes and Routers onboard Satellites", International Journal of Satellite Communications and Networking, vol. 25, no. 2, 2007, 195-216.

[11] R. Slywczak, "Low-Earth-Orbit Satellite Internet Protocol Communications Concept and Design", NASA/TM technical report, 2004-212299.

[12] T. Vladimirova, C. P. Bridges, G. Prassions, et al, "Characterising Wireless Sensor Motes for Space Applications", 2nd NASA/ESA Conference on Adaptive Hardware and Systems (AHS 2007), Edinburgh, UK, 5-8 Aug. 2007, pp. 43-50

[13] T. Vladimirova, X. Wu, C. P. Bridges, "Development of a Satellite Sensor Network for Future Space Missions", IEEE Aerospace Conference, Big Sky, MT, 1-8 March, 2008, pp. 1-10

[14] P. Dubois, C. Botteron, V. Mitev, et al, "Ad-hoc Wireless Sensor Networks for Exploration of Solar-System Bodies", Acta Astronautica, vol. 64, 2009, pp. 626-643

[15] P. Brenner, "A Technical Tutorial on the IEEE 802.11 Protocol", Breezecom Wireless Communications, 1997.

[16] A. J. Van Dierendonck, G. A. McGraw, R. J. Erlandson and R. Coker, "Cross-correlation of C/A codes in GPS/WAAS Receivers", ION GPS 99, Nashville, TN, 14-17 Sep. 1999.

[17] Z. Zhu and F. Van Graas, "Effects of Cross Correlation on High Performance C/A Code Tracking", ION NTM 2005, San Diego, CA, 24-26 Jan. 2005.

[18] E. Gill, O. Montenbruck, K. Arichandran, S. Tan and T. Bretschneider, "High-precision onboard orbit determination for small satellites-the GPS-based XNS on X-SAT", 6th Symposium on Small Satellites Systems and Services, La Rochelle, France, 20-24 Sep. 2004.

[19] L. P. Clare, J. L. Gao, E. H. Jennings and C. Okino, "A Network Architecture for Precision Formation Flying Using the IEEE 802.11 MAC Protocol", IEEE Aerospace Conference, Big Sky, MT, 5-12 March 2005.

[20] R. Sun, D. Maessen, J. Guo and E. Gill, "Enabling inter-satellite communication and ranging for small satellites", 9th Symposium on Small Satellites Systems and Services, Funchal, Portugal, 31 May – 4 June, 2010.

[21] J. Kim and B. Tapley, "Simulation of dual one-way ranging measurements", Journal of spacecraft and rockets, vol. 40, 2003, pp. 419-425.

[22] J. Spilker, "Signal structure and theoretical performance", Global Positioning System: Theory and Applications, American Institute of Aeronautics and Astronautics, Inc. vol. 1, 1996, pp. 57-105.

[23] P. Ward, J. Betz and C. Hegarty, "GPS satellite signal characteristics", Understanding GPS: Principles and Applications, 2nd ed., Artech House, Inc. 2006, pp. 135-145.

[24] M. Pratap and P. Enge, Global Positioning System: Signals, Measurements and Performance, 2nd ed., Ganga-Jamuna Press, 2006, pp. 365-379.

[25] K. Alfriend, S. R. Vadali, P. Gurfil, J. How and L. Breger, "Spacecraft Formation Flying – Dynamics, Control and Navigation", 1st ed., Elsevier, 2010, pp. 84-89.

Defining the Optimal Architecture for Multi-Standard Radio Receivers Embedding Analog Signal Conditioning

Silvian Spiridon[1,2], Claudius Dan[1] and Mircea Bodea[1]
[1]"POLITEHNICA" University of Bucharest, Romania
[2]Now with Broadcom, Bunnik, The Netherlands
e-mail: silvian.spiridon@gmail.com, claudius.dan@gmail.com, mirceabodea@yahoo.com

***Abstract*—This paper focuses on finding the optimal architecture for a multi-standard radio receiver embedding analog signal conditioning compatible with the major commercial wireless standards. By developing a standard independent methodology, the paper addresses systematically the large amount of information comprised in the envisaged standards. Based on the systematic approach, the multi-standard receiver main electrical requirements are defined and their values determined. The presented results constitute the starting point in building a multi-standard wireless receiver.**

Keywords-software defined radio; receiver electrical specifications.

I. INTRODUCTION

At the beginning of the mobile Internet age, it becomes clear there is a strong need of mobile equipment able to maximize its wireless connectivity.

There are several reasons that make extremely attractive the usage of multi-standard radio transceivers to enable the wireless interoperation of such mobile equipment, [1,2].

Firstly, a multi-standard solution is efficient as only one design is required to handle the mobile device wireless communication. Thus, the number of different dedicated ICs or IP blocks inside a large SoC is reduced. This simplifies the overall communication platform integration.

Secondly, since only one design is required to cover for all the targeted wireless standards all cost related to the IP development and testing are minimized.

The ideal multi-standard receiver was proposed by Mitola in [3]. The Software Defined Radio Receiver (SDRR) shown in Fig. 1.a is the optimal choice from system level perspective, as it comprises only an ADC.

In reality, due to practical implementation constrains, a multi-standard receiver requires a signal conditioning block in between the antenna and the ADC. The SDRR concept shown in Fig. 1.b relaxes the ADC specifications by ensuring additional selectivity and amplification for the wanted signal.

First, the paper focuses on finding the optimal architecture for the SDRR signal conditioning block that enables its monolithic integration.

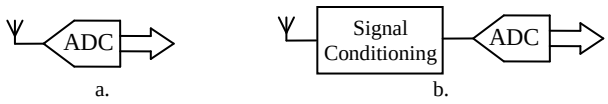


Figure 1. a. Mitola SDRR Concept, b. SDRR embeding signal conditioning.

Based on a detailed overview of the key receiver architectures, Section II introduces the optimal architecture for a SDRR that targets compatibility to the major commercial wireless standards listed in Table I.

Second, the paper pursues the identification of the SDRR key electrical requirements and their values. Based on a systematic approach, the paper introduces a standard independent methodology for evaluating the SDRR performance.

Section III defines the SDRR receiver sensitivity, NF and gain requirements, while Section IV analyses the blocker and interferers impact on the SDRR linearity requirements. Section V concludes the paper.

II. INTRODUCING THE OPTIMAL ARCHITECUTRE OF THE SIGNAL CONDITIONING BLOCK

This Section covers one of the paper's goals that is to find the optimal architecture for the SDRR signal conditioning block enabling its monolithic integration. In Section II.A, the key receiver architectures are analyzed, while Section II.B determines the quadrature direct conversion topology suits best the envisaged purpose.

Section II.C overviews the architectural changes that make the homodyne receivers ready for monolithical integration. The presented solutions are realized without introducing particular analog tricks to satisfy the needs of only one of the standards, as the SDRR must represent "universal receivers", and not be turned into a "multi-standard ASICs".

A. Overview of Key Receiver Architectures

First of all, one of the most popular receiver architectures is the heterodyne architecture. It was developed in 1918 by Edwin Armstrong as a viable alternative to the regenerative receiver with respect to the technical issues of vacuum tubes implementation, [4].

TABLE I. TARGETED MAJOR COMMERCIAL WIRELESS STANDARDS KEY SPECIFICATIONS

Wireless Standard		Frequency Plan [MHz]		Modulation Type	SNR_0 [dB]	RF Signal BW / Channel Spacing [MHz]	Specified Sensitivity [dBm]	Sensitivity @ NF_{RX} = 3 dB [dBm]
		Downlink	*Uplink*					
GSM, [5]	GSM 850	869 ÷ 894.8	824 … 849.8	GMSK	9	0.2	–102	–109
	GSM 900	935 ÷ 960	890 … 915					
	DCS 1800	1805 ÷ 1880	1710 … 1785					
	PCS 1900	1930 ÷ 1990	1850 … 1910					
UMTS, [6]	I	2110 ÷ 2170	1920 ÷ 1980	QPSK	−18 (@ 12.2kbps)	3.84 / 5	–117	–123
	II	1930 ÷ 1990	1850 ÷ 1910					
	III	1805 ÷ 1880	1710 ÷ 1785					
Bluetooth, [7]		2402 ÷ 2480		GFSK	16	1	–70	–94
DECT, [8]		1880 ÷ 1980, 2010 ÷ 2035		GFSK	13	1.2 / 1.736	–83	–97
WLAN IEEE 802.11b (DSSS), [9]	1 / 2 Mbit/s	2400 ÷ 2485		DBPSK / DQPSK	−4 / −2	14 / 5	–80	–104 / –102
	5.5 / 11Mbit/s			CCK	9 / 11		–76	–91 / –89
WLAN IEEE 802.11a,g (OFDM), [9]	6 / 9 Mbit/s	5150 ÷ 5350 & 5725 ÷ 5825 (a) 2400 ÷ 2485 (g)		BPSK	4 / 5	16.6 / 20	–82 / –81	–95 / –94
	12 / 18 Mbit/s			QPSK	7 / 9		–79 / –77	–92 / –90
	24 / 36 Mbit/s			16QAM	12 / 16		–74 / –70	–87 / –83
	48 / 54 Mbit/s			64QAM	20 / 21		–66 / –65	–79 / –78

The basic block schematic of this concept is depicted in Fig. 2. The original superheterodyne uses only one downconverter mixer (i. e., single conversion superheterodyne), and mixes the Radio Frequency, RF, input signal with the Local Oscillator, LO, signal.

The resulting signal frequency is mixed down by the mixer (i. e., MIX in Fig. 2), to an Intermediate Frequency, IF, equal to the difference between the RF carrier and LO signal frequencies.

Intrinsically the mixing process will render at the mixer output also the sum frequency component. For most applications this component represents an unwanted signal and is filtered by the band pass filter (i. e., BPF in Fig. 2) following the mixer and/or in the mixer output stage.

The major issue of the heterodyne topology is the image frequency rejection.

The issue is two symmetrical signals with frequencies spaced apart by twice the IF frequency are downconverted by LO mixing to the same IF frequency, as shown in Fig. 3.

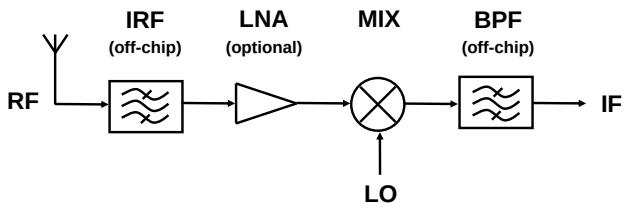


Figure 2. Single Conversion Superheterodyne Receiver Block Schematic.

Generally, only one of the two sideband signals conveys useful information, while the other represents an unwanted image signal.

Thus, for superheterodyne receivers image signal rejection is critical for a proper signal demodulation.

The superheterodyne architecture solves the issue by filtering the image signal before it enters the mixer, or more precisely, immediately after the antenna in the external image reject filter (i. e., IRF in Fig. 2).

The image rejection filter specifications depend on the IF value and they are more relaxed as the image frequency is larger, respectively as the distance between the RF carrier and its image is larger.

Signal conditioning constraints, set by the channel selection filter located after the down conversion mixer (i. e., the second band pass filter of Fig. 2), prevent the choice of a very large IF. This toughens image filtering requirements. In practice, ceramic filters satisfy the constraints, but they possess two major drawbacks: they are quite expensive and by far not compatible with monolithic integration.

Channel selection is also demanding, as for many applications channel bandwidth is fairly small compared with IF. In such context, bandpass Surface Acoustic Wave (SAW) filters are used for analog channel selection. But, these types of filters are unattractive to SoC ICs for the same two reasons as the ceramic antenna filters: incompatibility with monolithic integration and high cost.

Basically, single conversion superheterodyne receiver design is driven by the trade-off between antenna and channel filtering, which imposes the optimum IF frequency.

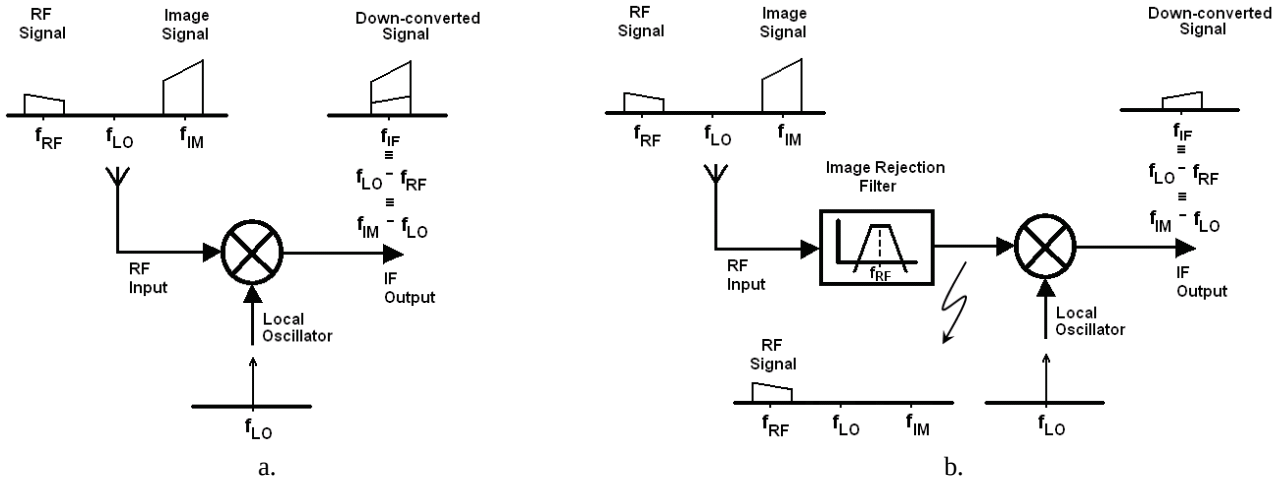


Figure 3. Down-conversion: a. without Image Rejection and b. with Image Rejection.

Second of all, so far, the design of superheterodyne receivers has been optimized to alleviate image rejection rather than optimizing RF performance. Thus, using an image rejection receiver that implements a complex mixer to cancel out the unwanted image signal, removes the lock on architecture enabling the RF performance optimization. The principle schematic of such a receiver is depicted in Fig. 4.a.

The complex mixer is made out of two mixers which share the same RF input, while the LO port is controlled by two quadrature signals. By adding a 90 degrees delay line in one signal path, the down converted image signals will be in-phase, while the useful signals will be 180 degrees delayed. Hence, by considering the difference between the two paths the image signal is cancelled, while the useful signal is added. The major advantage of this approach is the antenna filtering becomes less critical. Thus, the use of expensive and bulky, external (off-chip), ceramic filters are no longer required. On the other hand, the image rejection now depends on the quadrature accuracy of both gain and phase of the LO and IF paths. If the two LO signals exhibit exactly 90 degrees phase delay and have the same amplitude, while the gain of the two paths are perfectly matched, the unwanted image signal is completely rejected.

Hence, image rejection receivers cancel out the image signals by subtracting two – potentially very large – signals, resulting in a difference that is theoretically equal to zero.

However, any gain error or phase error between the two signal paths will determine an incomplete annihilation of the image signal. Thus the image rejection, *IR*, is given by, [10]:

$$IR = -20\log\left\{\frac{1}{2}\left[\frac{Gain_{I-Q\,err}}{Gain} + tg\left(Phase_{I-Q\,err}\right)\right]\right\} \quad (1)$$

where *Gain* represents the receiver's gain, $Gain_{I\text{-}Qerr}$ is the I-Q gain mismatch and $Phase_{I\text{-}Qerr}$ is the I-Q phase mismatch.

Since accurate wide-band quadrature phase shifters are difficult to design, Weaver receivers of Fig. 4.b are preferred.

To cancel the need for the 90 degrees phase shift on the signal path an extra pair of mixers and quadrature LO signals are required. Still, the LO signals quadrature accuracy, of both gain and phase, and the gain matching of the quadrature downconverted channels set the image rejection performance as described by (1).

Third of all, all receiver architectures presented so far have to fight image rejection. In direct conversion receivers also known as homodyne receivers, the IF frequency is zero, hence, the useful signal is its own image.

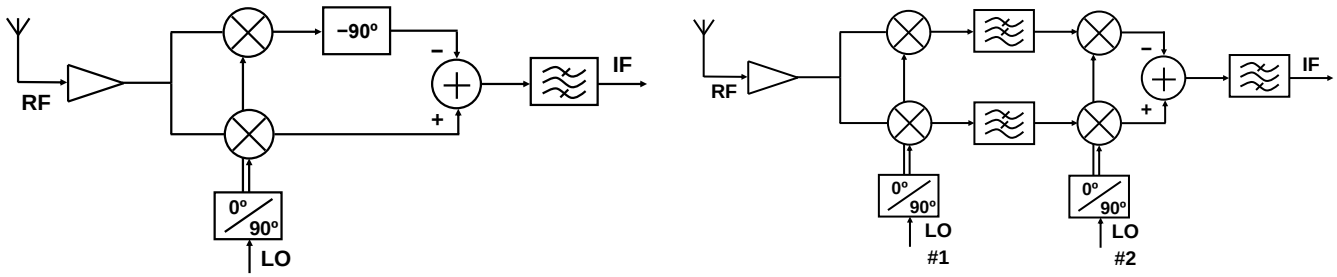


Figure 4. a. Image Rejection Receiver and b. Weaver Receiver.

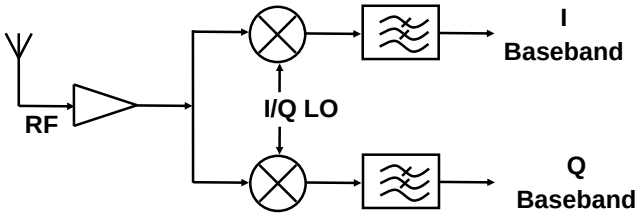


Figure 5. Quadrature Direct Conversion Receiver Block Schematic.

Therefore, in a homodyne receiver (see Fig. 5) the image signal has an amplitude comparable to the useful signal, and hence, image rejection requirements are relaxed. Furthermore, all baseband processing, like analogue filtering, analog-to-digital conversion and the digital demodulation, take place at the lowest possible frequency.

These features make the homodyne receiver an ideal candidate for monolithic integration and open the possibility of creating an "*universal*" receiver, compatible with all wireless standards, [4].

However, although direct conversion receivers monolithic integration seems straight forward there are several drawbacks to this approach.

The second major issue of direct conversion architecture is that even order distortions generate a signal dependent DC offset.

Another issue of such architecture is self-mixing. Basically, the LO signal, which in most cases is orders of magnitude larger than the RF signal, leaks to the RF port of the mixer and it will be mixed down to baseband. If the LO leaking signal is phase shifted with respect to the real LO, this almost always being the case in practice, the DC offset caused by self-mixing may dominate the mixer output.

Finally, although direct conversion architecture has relaxed image rejection specifications, it has to fight with DC offset, 1/f noise and self-mixing.

Hence, the low-IF architectures (see Fig. 6) become attractive. Essentially, the RF signal will now be downconverted to a low IF frequency (i. e., up to a few hundred kHz) and thus, the issues of direct conversion receivers are alleviated.

But, the image rejection requirements are again tight. This stresses the implementation of the active poly-phase filter that follows the mixer and is used for image rejection and channel selection.

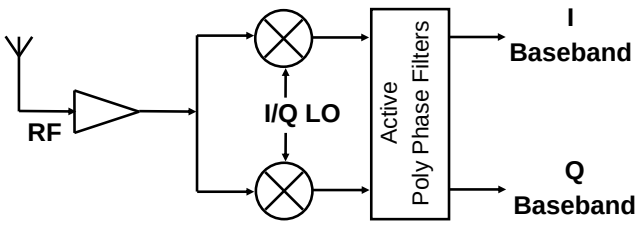


Figure 6. Low-IF Conversion Receiver Block Schematic.

B. SDRR Architecture Choice: Heterodyne vs. Homodyne

The main features of a SDRR must be a versatile architecture and the ability to be reconfigured on-the-fly as the communication burst requires.

From the perspective of SoCs, the optimization of the chip power dissipation and die area is mandatory. As the SDRR will be a part of a SoC, this trade-off must be the main guideline in sizing the SDRR design, as well as in choosing its architecture, as a first and, very important, starting point.

From area perspective the cumbersome image rejection filters of superheterodyne topology are not so attractive for monolithic integration.

On the other hand, for direct conversion architecture the image rejection requirements are much smaller than for any other receiver architecture.

Furthermore, the IF selection for superheterodyne architectures is fairly cumbersome and cannot be extrapolated in a systematic way to all standards, as it would be required for a true SDRR, [11].

The IF should be chosen to avoid the in-band down-conversion of strong interferers. In most applications the nearest strong interferers is located three channels apart from the RF carrier, [12]. As the channel bandwidth differs even within the same wireless standard, it is not possible to select intermediate frequencies which will lead to reuse of same image filters for a multi-standard environment compatible receiver.

Similarly, it is difficult to design an accurate wide-band 90° phase shift block for the image rejection receiver of Fig. 4.a, as imposed by a multi-standard application. Also, the fact the Weaver receiver from Fig. 4.b requires a cascade of two complex mixers increases the overall receiver area and power consumption, while it is also constraining its image rejection capabilities.

From power consumption perspective the homodyne topology has even more advantages. First of all, the baseband signal processing takes place at the lowest possible frequency. Secondly, this topology is not tributary to the 3 dB noise penalty of superheterodyne architectures, [11].

Hence, the SDRR architectures of choice are superheterodyne, homodyne or low-IF.

Table II summarizes the advantages and disadvantages of the three architectures with respect to monolithic integration in a SDR SoC.

Given the overview presented in Table II, it becomes clear the direct conversion receivers represent the optimal choice for satisfying the requirements of a true SDRR.

This has been validated through several circuit implementations in CMOS processes, [1, 13-15].

As noted, zero-IF receivers are susceptible to multiple issues (e. g., DC offset, 1/f noise, the self-mixing process), which make their monolithic integration quite a challenging task. All these aspects will be discussed in the following sub-section.

TABLE II. SUPERHETERODYNE, HOMODYNE AND LOW-IF RECEIVERS FOR SOC INTEGRATION – PROS AND CONS

Superheterodyne		Homodyne		Low-IF	
PROs	CONs	PROs	CONs	PROs	CONs
☺ Well known	☹ Requires off-chip components Ceramic antenna filter SAW Filter	☺ Less external components No ceramic antenna filter No SAW Filter	☹ DC Offset	☺ no DC Offset	☹ High image rejection requirement
			☹ 1/f noise	☺ reduced 1/f noise	
	☹ IF selection Difficult to mitigate the multi-standard environment	☺ Image is wanted signal mirror Mirror signal is not a strong interferer	☹ self-mixing	☺ reduced self-mixing	
	☹ Power consumption BB signal conditioning is done at IF	☺ Power consumption BB signal conditioning is done at lowest frequency			
	☹ 3 dB noise penalty, [10] Image frequency band degrades receiver SNR by 3 dB	☺ No 3 dB noise penalty Quadrature receiver			

C. *Making the Homodyne Architecture Ready for Monolithical Integration*

As detailed in the previous sub-section and in-depth analyzed in [16], due to intrinsic operation of homodyne systems, they exhibit a large sensitivity to DC offset, either static or dynamic, and 1/*f* noise. Also, self-mixing issues can dramatically reduce performance of receivers implemented with direct conversion architectures.

First of all, homodyne architecture is extremely sensitive to static DC offsets and 1/*f* noise. As for some wireless standards a large part of the signal energy is found at low frequencies (e. g., GSM), the down-converter mixer output is DC coupled to the anti-alias LPF. Thus, the receiver output risks of being overloaded even for small values of the DC offset, in the order of a few hundreds μV given the large VGA gain, usually larger than 60 dB, [17]. Regular AC coupling will not solve the issue, as receiver settling will be severely affected by a low cut−off frequency in the order of a few hundred Hz.

Hence, the receiver must embed an offset calibration loop.

One possibility is to use the correlated double sampling offset compensation technique, [18]. This is preferred to chopper stabilization, [19], as there is no risk of spurs overwhelming the receiver output spectrum. The receiver block schematic embedding offset calibration is shown in Fig. 7, redrawn from [16].

In the first phase (i. e., *Offset_meas* control signal is "High" – switches closed) the baseband chain DC offset is sampled on a capacitor, via the additional transimpedance amplifier, while the antenna input is shorted to ground, [10].

During normal operation, the second phase (i. e., *Offset_meas* is "Low" – switches open), the RF input is connected back to the antenna and the signal flows through the receiver, while the DC offset is inherently cancelled out.

The static DC offset compensation loop will also reduce the 1/f noise level.

The Fig. 7 receiver embeds multiple LNAs to mitigate the multi-standard frequency plan requirements from Table I.

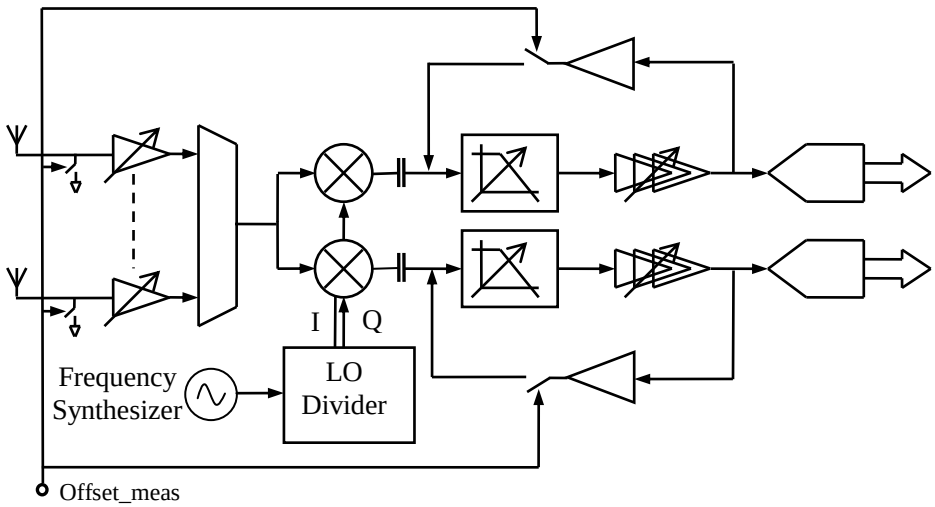


Figure 7. Low-IF Conversion Receiver Block Schematic.

The second major issue of the homodyne receiver architecture is the signal dependent DC offset generation due to even order distortions. the received input power can change dynamically, since other transmitters may start to communicate, a dynamic offset component is generated due to the receiver second order non-linearity. Therefore, dynamic offset compensation, to the extent required by almost all commercial wireless standards, implies the usage of a differential architecture for the whole receiver chain, starting with its LNA.

Thirdly, the self mixing process, determined by the LO mixing with the LO signal leaking the VCO to the receiver input, can generate a large DC offset overloading the receiver output. Hence, the VCO must not oscillate at the same frequency with the RF carrier frequency. Hence, the quadrature LO signals driving the downconverter mixer must be obtained by dividing down the VCO frequency in the LO divider block from Fig. 7 (e. g., [20]). Moreover, very good isolation between RF and LO mixer ports is required for good receiver performance.

III. Defining the SDR Sensitivity, Noise Figure and Gain Requirements

A. Introducing the Minimum Signal-to-Noise Ratio Required for Proper Signal Demodulation, SNR_0

The front-end must be able to downconvert the useful signal without hampering its electrical properties, such as the baseband processor is able to demodulate the information within a specified Bit Error Rate (*BER*). Basically there is a minimum SNR at the receiver output, SNR_{out}, required for the digital demodulator to properly demodulate the useful signal. This minimum SNR_{out} value is further on denoted as SNR_0.

As expected, higher SNRs are required to demodulate the signal within the same BER as the modulation number of bits per symbol increases.

Increasing the SNR requirements is achieved at the cost of higher power consumption, by increasing the signal power, or at the cost of lowering the bandwidth. In any case, there is a trade-off between power consumption and BitRate (*BR*). This can be mathematically expressed as follows:

$$\begin{cases} S = BR \cdot E_b \\ N = BW \cdot N_0 \end{cases}, \quad (2)$$

where *S*, respectively *N*, is the receiver input signal, respectively noise, power, E_b is the energy per bit, N_0 is the noise power density at the receiver input; in practice, $N_0 = k_B T \cdot F$ – where F is the receiver noise factor.

Given (2), the maximum bit-rate from can be written as, [2]:

$$BR \le BW \log_2\left(1+\frac{S}{N}\right) = BW \log_2\left(1+\frac{BR \cdot E_b}{BW \cdot N_0}\right) \quad (3)$$

From (3) results there is a minimum amount of signal energy required to transmit a bit:

$$E_b \ge N_0 \frac{2^{BR/BW}-1}{BR/BW} \quad (4)$$

As (4) shows, the minimum E_b only depends on N_0 and on the coding scheme, through the *BR*/*BW* ratio.

There are two extreme cases depending on the *BR*/*BW* ratio value.

Firstly, if *BR*/*BW* is very low (i. e., a large *BW* is used for a small *BR*) the limit from (4) is:

$$E_b \ge N_0 \ln 2 \quad (5)$$

This case is exploited by spread spectrum systems. Basically, the SNR_0 has a negative value for UMTS and WLAN 802-11 standards, as it accounts the processing gain, [21].

Secondly, for large *BR*/*BW* (e. g., for 64QAM), the limit from (4) becomes:

$$E_b \ge N_0 \frac{2^{BR/BW}}{BR/BW} \quad (6)$$

Hence, the SNR_0 in this case is:

$$SNR_0 = \frac{BR \cdot E_b}{BW \cdot N_0} \ge 2^{BR/BW} \quad (7)$$

As an example the SNR from (7) translates to 18 dB SNR_0 for 64QAM. Of course, this theoretical limit translates to a few dB higher value in practical implementation.

Based on the analysis presented in [22], the SNR_0 as a function of the *BER* has been determined for the basic modulation schemes. Table I notes the targeted standards signal modulation and the corresponding SNR_0 values.

The usage of the concept "SNR_0" facilitates the finding of the multi-standard receiver key electrical parameters by enabling a standard independent approach.

B. Sensitivity and Noise Figure

One of the most important parameters of a wireless receiver is its sensitivity, S_{RX}. The sensitivity is defined as the minimum input signal the receiver must be able to demodulate within the specified *BER*.

Thus, the Signal-to-Noise Ratio at the RX output, SNR_{out}, has to be above SNR_0. As each standard specifies a sensitivity level, given the useful signal RF bandwidth, BW_{RF}, the receiver NF, NF_{RX}, is calculated as:

$$NF_{RX} \le S_{RX} - 10\log BW_{RF} - SNR_0 - N_0, \quad (8)$$

where $N_0 = k_B T = -174$ dBm/Hz represents the noise power spectral density at the antenna output for T = 290 °K.

In practice, an overhead to SNR_0 should be considered in (8), since the overall receiver SNR is degraded by multiple factors, not only by noise (e. g., imperfect impedance matching, multipath channel).

The receiver NF specifications for all the wireless standards can be calculated with (8) by accounting the specified sensitivity levels from Table I.

A NF as the one derived by (8) can be obtained at the expense of larger power consumption of thy receiver.

In order to maximize the link budget, most commercially available dedicated receivers push their sensitivity level towards smaller and smaller values by decreasing NF_{RX}.

Hence, a true re-configurable multi-standard solution must embed a receiver with a small NF (typically < 3 dB) in order to be able to achieve a low enough sensitivity for all the targeted standards. Table I also comprises the required sensitivity levels, assuming $NF_{RX} = 3$ dB, for all the envisaged wireless standards.

C. Maximum gain requirements

In general, the receiver signal conditioning path gain is constraint by the received signal strength.

Any wireless receiver with an analog signal conditioning path embeds at least one variable gain block.

Thus, for each communication burst an Automated Gain Control loop (AGC) measures the Receiver Signal Strength Indicator (RSSI) and changes the receiver gain accordingly, in order to avoid the ADC overloading and to optimally load it.

The receiver maximum gain requirements are constraint by the ADC full scale level, FS_{ADC}, and the specified receiver sensitivity.

In order to optimally load the ADC, the receiver signal conditioning path maximum gain, G_{RX}, is given by:

$$G_{RX} = \frac{k \cdot FS_{ADC}}{S_{RX}}, \quad (9)$$

where $k < 1$ accounts the head room taken to avoid the ADC overloading.

TABLE III. THE MULTI-STANDARD RECEIVER MAXIMUM GAIN REQUIREMENTS

Wireless Standard		G_{RX} [dB]
GSM		115
UMTS		129
Bluetooth		100
DECT		103
WLAN IEEE 802.11 b,g (DSSS)	1 / 2 / 5.5 / 11 Mbit/s	110 / 108 / 97 / 95
WLAN IEEE 802.11 a,g (OFDM)	6 / 9 / 12 /18 24 /36 /48 /54 Mbit/s	101 / 100 / 98 / 96 93 / 89 / 85 / 84

For a multi-standard receiver embedding analog signal conditioning (see Fig. 1.c), k is set by the Variable Gain Amplifier (VGA) gain step (e. g., 6 dB, [14]). Equation (9) assumes all interferes and blockers have been totally filtered out before the ADC. This corresponds to the case of complete analog channel selection.

Table III presents the receiver signal conditioning path maximum gain requirement for all the envisaged standards calculated based on equation (9) for a 1 V FS ADC and for a receiver matched to 100 Ω with k = 0.9; the NF_{RX} of the receiver was assumed to be 3 dB. For direct sequence spread spectrum systems (e. g., WLAN 802.11b) the gain requirement is smaller, as in practice other signals will be present as well inside the received bandwidth.

Nonetheless, we can conclude, based on Table II data, that the low sensitivity levels of the targeted wireless standards require a large maximum gain for the multi-standard receiver.

IV. BLOCKERS AND INTERFERERS IMPACT ON THE MULTI-STANDARD RECEIVER CHARACTERISTICS

Besides the useful signal, other interferers and blockers can be present at the antenna input. The list comprising all interferes and blockers under which influence the receiver must still be able to properly demodulate the wanted signal represents the receiver blocker diagram. For each wireless standard such a receiver blocker diagram is specified.

Based on the envisaged wireless standards blocker diagram analysis, in [12] a receiver generic blockers diagram has been constructed. This newly introduced tool enables the designer to handle efficiently the large amount of information comprised in the targeted radio standards.

Thus, based on the receiver generic blockers diagram analysis it results immediately there are two major issues due to blockers and interferes:

- the receiver output clipping, due to the large receiver gain requirements and to the large difference between the useful signal and the blocker levels (i. e., typically > +40 dBc);
- intermodulation distortions that fall in-band, due to the receiver not perfectly linear transfer characteristic.

The receiver output clipping can be handled (i) by making the LNA and VGA blocks gain variable and (ii) by allowing the receiver high frequency part noise and linearity performance to adjust with the RF front-end gain. This has analysed in-depth in [23].

On the other hand intermodulation distortions are un-wanted products that potentially fall in-band and cannot be disseminated from the useful signal. Thus the wanted signal demodulation is affected due to the SNR degradation. Further on the analysis presented in this Section focuses on finding the values for the Figures of Merit (FOMs) used in evaluating the radio receiver linearity performance: the *IIP*2 and *IIP*3.

A. Finding the SDR IIP2

While receiving the RF input power may change significantly because of the reception of unwanted blockers/interferers. Due to the receiver even order

distortions, the received signal DC offset component will change. This *dynamic offset* effect upsets the received signal demodulation, especially if the envisaged modulation concentrates a large part of the symbol spectral power at low frequencies. This is the case for older standards like GSM, as the latest wireless standards use modulation schemes that do not carry information at low frequencies.

The figure of merit quantizing the analog front-end second order distortions is the second order intercept point, *IP*2. The SDRR input referred IP2, $IIP2_{RX}$, is given by, [10]:

$$iIP2_{RX} = 2 \times P_{blocker} - P_{in} + SNR_0, \quad (10)$$

where $P_{blocker}$ is the blocker level and P_{in} is the wanted signal level.

Based on the targeted standards analysis, it results the worst case scenario is met for the GSM standard that requires $IIP2_{RX}$ = +46 dBm, [2].

B. Finding the SDR IIP3

For most wireless receivers, given the fully differential circuit implementation, the dominant non-linear contribution comes from the third order coefficient of power series expansion of their transfer characteristic. The maximum in-band level of the third-order intermodulation product, P_{IM3}, must be smaller than the useful RF signal level with SNR_0:

$$P_{IM3} \le P_{in} - SNR_0 \quad (11)$$

In practice, a supplementary head room to SNR_0 should be considered, since the overall receiver SNR is degraded by multiple factors, not only by the down-converted spurs.

TABLE IV. MULTI-STANDARD RECEIVER IIP3 REQUIREMENTS

Standard	Intermodulation conditions	RX IIP3[dBm]	
		Eq.	*Value*
GSM	$P_{interferer}$ @ −49 dBm, P_{in} @ −99 dBm	(12)	−19
UMTS	$P_{interferer}$ @ −46 dBm, P_{in} @ −114 dBm	(12)	−21
Bluetooth	$P_{interferer}$ @ −39 dBm, P_{in} @ −64 dBm	(12)	−18.5
DECT	$P_{interferer}$ @ −47 dBm, P_{in} @ −80 dBm	(12)	−24
WLAN IEEE 802.11b,g (DSSS)	$P_{interferer}$ @ −35 dBm, P_{in} @ −70 dBm CCK – 11 Mbit/s	(12)	−12
W-LAN IEEE 802.11g (OFDM @ 2.4 GHz)	Interferer intermodulation: $P_{interferer}$ @ Sensitivity, P_{in} @ +32…+15 dBc (6…54 Mbit/s)	(12)	−32
	Blocker intermodulation: $P_{blocker}$ @ −10 dBm, P_{in} @ −42 dBm, BPSK – 6 Mbit/s	(12)	+8.5
	Sub-carrier intermodulation: P_{in} @ −20 dBm, N = 52 carriers, 64QAM – 54 Mbit/s	(15)	+10
W-LAN IEEE 802.11a (OFDM @ 5 GHz)	Sub-carrier intermodulation: P_{in} @ −30 dBm, N = 52 carriers, 64QAM – 54 Mbit/s	(15)	+5

Given (11), the receiver IIP3, $IIP3_{RX}$, must meet the condition specified by the equation:

$$IIP3_{RX} \ge P_{interferer} + \frac{P_{interferer} - P_{IM3}}{2} \quad (12)$$

where $P_{interferer}$ is the power per interferer of two interferers that cause the in-band third order distortion.

A special case is represented by OFDM Signals. An OFDM signal comprises frequency orthogonal sub-carriers, [19]. Receiver non-linearity leads to formation of bogus signals in-band due to sub-carrier intermodulation. The figure of merit in evaluating the third order intermodulation products thus formed is the Composite Triple Beat (CTB).

As is pointed out in [24] the worst case for the CTB products level is found in the centre band of the OFDM signal spectrum:

$$CTB[\text{dB}] \le -2(IIP3_{RX} - P_{in}) + 1.74, \quad (13)$$

where P_{in} is the OFDM signal power in all the carriers.

Hence, in order for the digital back-end to be able to still demodulate properly the wanted signal, the CTB level must be smaller than the useful RF signal level per carrier with SNR_0:

$$CTB \le P_{in} - 10\log N - SNR_0, \quad (14)$$

where *N* represents the number of OFDM sub-carriers.

In (14) SNR_0 represents the corresponding SNR headroom of the OFDM sub-carrier modulation.

Given (13) and (14), it results that in order to avoid destructive inter-carrier intermodulation, the $IIP3_{RX}$ must meet the following condition:

$$IIP3_{RX} \ge \frac{1}{2}(P_{in} + 10\log N + SNR_0 + 1.74) \quad (15)$$

Each wireless standard specifies a set of particular intermodulation conditions. Table IV summarises the power per interferer of two interferers that cause the in-band distortion and the input signal power. By analysing all the targeted standards, the receiver *IIP*3 specifications were derived using (12) or (15) and noted in Table IV. The large variations in the *IIP*3 requirements are a reflection of the extreme reception conditions specific to the wireless environment. In [23] it is shown a versatile receiver is able to mitigate all presented scenarios, by adjusted dynamically its linearity and noise performance with the received power.

V. CONCLUSIONS AND FUTURE WORK

This paper conducted an analysis for finding the optimal architecture for a SDRR embedding analog signal conditioning and targeting compatibility with the major commercial wireless standards (see Table I).

Based on the analysis of the key receiver architectures, the homodyne receiver is found to suit best a monolithic implementation of a SDRR. By enhancing the classical homodyne architecture through the inclusion of an offset cancelation loop and quadrature LO generators the SDRR monolithic integration is made possible. The presented solutions are realized without introducing particular analog tricks to satisfy the needs of only one of the standards, as the SDRR must represent "universal receivers", and not be turned into "multi-standard ASICs".

Further on, the minimum SNR at the receiver output, SNR_0, required for a proper signal demodulation was calculated for all the envisaged standards. The usage of the concept "SNR_0" facilitates the finding of the multi-standard receiver key electrical parameters by enabling a standard independent approach.

Thanks to the standard independent systematic approach the presented analysis found the values for the key SDRR electrical specifications (i. e., NF_{RX}, $IIP2_{RX}$ and $IIP3_{RX}$) that ensure its compatibility with the envisaged standards.

Of course, a true SDRR has to be versatile and robust, such as it can adjust dynamically its performance (e. g., NF_{RX}, $IIP3_{RX}$) depending on the communication burst particularities. Nonetheless if a SDRR targets compatibility with the standards from Table I, it must meet the electrical specifications determined in this analysis.

So, the presented analysis constitutes the starting point in building the SDRR. Further on, the SDRR electrical specifications must be partitioned over its building blocks. The optimal specification partitioning must account the limitations due to the physical implementation (i. e., CMOS process) for each of the SDRR building blocks.

ACKNOWLEDGMENT

The authors would like to express their acknowledgment to Dr. F. Op't Eynde for the fruitful discussions on the topic.

REFERENCES

[1] S. Spiridon et. al., "Deriving the Key Electrical Specifications for a Multi-standard Radio Receiver," Proc. Of the First Intl. Conf. on Advances in Cognitive Radio COCORA 2011, April 2011, pp. 60-63.

[2] S. Spiridon, *Analysis and Design of Monolithic CMOS Sofware Defined Radio Receivers*, PhD Thesis, Ed. Tehnica, 2011.

[3] J Mitola, "Software radios – survey, critical evaluation and future directions," IEEE Nat. Telesystems Conf., 1992, pp. 13/15 – 13/23.

[4] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, 2nd Ed., 2004, pp. 710-713.

[5] "ETS 300 577, GSM: Digital Cellular Telcommunications System (Phase 2); Radio Transmission and Reception," European Telecommunication Standard Institute (ETSI), 1997.

[6] "3GPP, Technical Specification 25.101, v.6.10.0," Third Generation Partnership Project, Dec. 2005.

[7] "IEEE Std 802.15.1™–2005," IEEE Computer Society, 2005

[8] "ETS 300 175, DECT: Digital Enhanced Cordless Telecommunications," European Telecommunication Standard Institute (ETSI), 1995.

[9] "IEEE Std 802.11g−2003," IEEE Computer Society, 2003.

[10] F. Op't Eynde, "Direct-Conversion Radio Transceivers," RF IC Design Course Slides, EPFL, Switzerland, October 2005.

[11] F. Op't Eynde, "Front-end circuit design for RF Transceivers, "Cellular and WLAN Transceivers: From Systems to Circuit Design," Short Course Slides, ISSCC 2011, February 2011.

[12] S. Spiridon, C. Dan, M. Bodea, "Filter partitioning optimum strategy in homodyne multi-standard radio receivers," Proc. of the 7th Conf. on Ph.D. Research in Microelectronics and Electronics, PRIME 2011, July 2011, pp.9-13.

[13] J. Craninckx et. al, "A Fully Reconfigurable Software-Defined Radio Transceiver in 0.13µm CMOS," *Digest of Technical Papers of the International Solid State Circuit Conference*, ISSCC 2007, pp. 346-347 and 607.

[14] V. Giannini et. al, "A 2mm^2 0.1-to-5GHz SDR Receiver in 45nm Digital CMOS," *Digest of Technical Papers of the International Solid State Circuit Conference*, ISSCC 2009, pp. 408-409.

[15] M. Ingels et. al, A 5mm^2 40nm LP CMOS 0.1-to-3GHz multistandard transceiver, *Digest of Technical Papers of the International Solid State Circuit Conference*, ISSCC 2010, pp. 458-459.

[16] S. Spiridon et. al., "Making Homodyne Receivers Ready for Monolithic Integration in Multi-Standard Wireless Transceivers," Annals of Academy of Romanian Scientist, Series on Science and Technology of Information, Vol. 3, No. 2, pp. 73-80, 2010.

[17] S. Spiridon and F. Op't Eynde, "Low power CMOS fully differential variable-gain amplifier," in Proc. of the Annual Intl. Semiconductor Conf. CAS 2005, October 2005, vol. 2, pp 383-386.

[18] B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill, 2000.

[19] A. Danchiv, M. Bodea, C. Dan, "Amplifiers Chopper Technique Effects on Noise Performances," Proc. 10th Intl. Conf. on Optimization Of Electrical And Electronic Equipment OPTIM 2006, May 2006, pp. 16-20.

[20] S. Spiridon, F. Agavriloaie, M. Serban, "High Frequency Programmable Wide-Band Frequency Divider Design for CMOS Software Defined Radio Transceivers," Proc. of the 30th Annual Intl. Semiconductor Conf., CAS 2007, October 2007, Vol. 2, pp. 451-454.

[21] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, 1995.

[22] A. Tarniceriu, B. Iordache, and S. Spiridon, "An Analysis on Digital Modulation Techniques for Software Defined Radio Applications," Proc. of the 30th Annual Intl. Semiconductor Conf. CAS 2007, October 2007, vol. 2, pp. 571-574.

[23] S. Spiridon et. al, "Smart gain partitioning for noise – linearity trade-off optimization in multi-standard radio receivers," *Proceedings of the 18th International Conference Mixed Design of Integrated Circuits and Systems*, MIXDES 2011, June 2011, pp. 466-469.

[24] The Relationship Of Intercept Points Composite Distortions And Noise Power Ratios, http://www.matrixtest.com/literat/mtn109.pdf, Matrix Technical Notes, October 2005

Optimizing the Performance of a Dismounted Future Force Warrior by Means of Improved Situational Awareness

Tapio Saarelainen
Department of Military Technology
National Defence University
Finland
e-mail: tapio.saarelainen@mil.fi

***Abstract*—The future operational settings involve a battlespace where warriors and commanders rely on SA-tools to perform optimally in their given tasks. This may include combat settings in Military Environment (ME) as well as counter insurgency actions, peace-keeping operations and operations in Crises Management Environment (CME). In multi-national operations taking place in versatile and hostile environments, it is essential to detect, classify and identify the encountered objects and targets in the battlespace early enough. Consequently, the concept of war has changed in the direction of multi-symmetric warfare involving enemy troops, own forces and impartial entities. This paper discusses the existing and applicable Commercial-off-the-shelf (COTS)-based communication technologies with solutions suitable for military operations. These examples are examined by focusing on enhanced Situational Awareness (SA) as a tool facilitating improved decision making processes to support the execution of operations in a versatile battlespace and thereby optimize the performance of a dismounted Future Force Warrior (FFW).**

Keywords-Situational Awareness (SA), Common Operational Picture (COP), Wireless Polling Sensor Network (WPSN), Performance, Future Force Warrior (FFW).

I. INTRODUCTION AND DEFINITIONS

The purpose of this paper is to examine the existing and applicable Commercial-off-the-shelf (COTS) based communication tools available for enhancing dismounted Future Force Warriors' Situational Awareness (SA) and thereby maximizing their performance via improved decision-making capabilities. Dismounted warrior fights on foot, not inside of a vehicle. This paper begins by introducing the topic and key terminology followed by covering related studies [1]. Then the text turns to challenges concerning Combat Identification (CID). Then a comprehensive approach to targeting process is discussed. After this, the paper examines the means of accruing data for improved decision making processes and moves on to challenges involved in distributing Situational Awareness-data. Then the focus shifts to location and communication possibilities in urban areas followed by the discussion section and the conclusions.

When optimizing the performance of Future Force Warriors (FFWs), the data distributing and processing capabilities become seminal in enhancing improved SA and data distribution to enable near real-time Common Operational Picture (COP) and Shared Situational Awareness (SSA). Once the location data of varying entities can be reliably forwarded to respective command posts, the number of fratricide incidents and collateral damage can be significantly minimized.

Once identifying and defining relevant information and its distribution in the battlespace is determined as the key in Network Centric Operations (NCO), every effort to ensure the information flow between own warriors and sensors needs to be analyzed [2]. Contemporary weapon systems require greater amounts of intelligence data at a higher fidelity than ever before [3]. Since operations tend to be multi-national, different sensors and systems are required to communicate understandably between each entity to minimize fratricide and collateral damage by maximizing the distribution of the near real-time COP. One solution involves utilizing Business Management Language (BML) [2].

This paper concentrates on tackling the following three questions: 1) How to optimize the performance of a dismounted FFW by means of improved SA? 2) How to increase SA with the available COTS-based communication technologies? 3) What are the means to avoid casualties, collateral damage, and fratricide?

As for key terminology, a new network structure called the Wireless Polling Sensor Network (WPSN) is explained in [4]. Since nodes do not form a network per se but rather are polled by a selected node of the mobile network, they remain undetected due to their passive nature. The network structure offers a new and ubiquitous way to share and forward all kinds of data, including data collected by various sensors. Moreover, the outdated Identification Friend or Foe (IFF) systems are replaced and supplemented with effective and accurate means to identify the prevailing objects.

Examining the means to minimize fratricide and collateral damage presupposes applying the model presented in Figure 1. This terminologically updated model emphasizes how Tactics Techniques and Procedures (TTP), Combat Identification (CID), Common Operational Picture (COP), and Situational Awareness (SA) play a central role in minimizing the number of fratricide incidents and collateral damage.

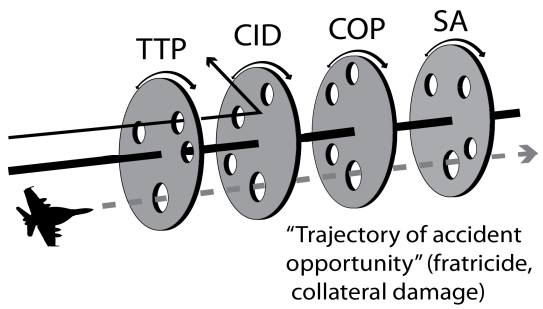


Figure 1. The Reason's Swiss Cheese Model updated by applicable terminology as a tool to explain the mechanism of avoiding fratricide and collateral damage, Blue-on-Blue (BoB) [5].

An applicable definition for Situational Awareness (SA) is given in the Army Field Manual 1-02 (September 2004): "Knowledge and understanding of the current situation which promotes timely, relevant and accurate assessment of friendly, competitive and other operations within the battlespace in order to facilitate decision making. An informational perspective and skill that fosters an ability to determine quickly the context and relevance of events that is unfolding."

The process of determining the affiliation of detected objects in the battlespace equals Target Identification (TI) [6]. When using this categorization, blue denotes the friendly force, red the enemy, and white refers to neutral (impartial) entities. The traditional method of TI is based on visual signature of the object of interest. In contemporary warfare TI is also based on utilizing the electromagnetic spectrum of the target. Properly applied, data and sensor fusion can be seen as a means to prevent collateral damage and fratricide. As a matter of fact, TI can be divided into two categories: Cooperative Target Identification (CTI) and Non-Cooperative Target Identification (NCTI). CTI allows a human shooter or sensor to interrogate a potential target and thereby forces the potential target to respond to the interrogation in a timely manner as described in Figure 2 [5] [6].

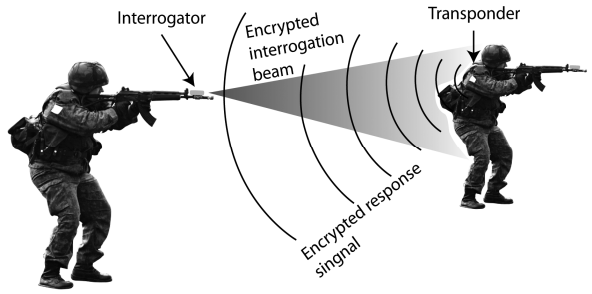


Figure 2. The process of Cooperative Target Identification (CTI) [5].

In contrast, NCTI does not require a cooperative response from the target. NCTI involves systems or methods which exploit the physical characteristics of entities in the battlespace to help identify and determine affiliation. NCTI systems include optics, such as Thermal Weapon Sights (TWS), night Vision Goggles (NVG), Forward Looking Infrared Radar (FLIR), as well as vehicle and personnel markings, for instance, Joint Combat Identification Marking Systems (JCIMS) [6].

Combat Identification can be defined as a process of attaining an accurate and timely characterization of detected objects in the joint battlespace to the extent that high confidence, timely application of military options and weapons resources can occur [6][8]. An extension of this can be understood as a process of accurately characterizing the detected objects via the operational environment sufficiently to support engagement decisions [6]. The purpose of CID is to enhance unit combat effectiveness and simultaneously minimizing fratricide. In the form of an equation CID reads as: SA + TI = CID [6].

The core capability in SA is Common Operational Picture that fosters effective decision making, rapid staff actions, and appropriate mission execution [6][9]. COP is employed to collect, share and display multi-dimensional information to facilitate collaborative planning and response to security incidents. COP typically comprises three types of modules as indicated in [5]: 1) information gathering sources that observe events and report information to the command and control module, 2) a command and control module that makes decisions based on both information received directly from its information gathering sources and information reported by other peers, and 3) display units at the emergency location that receive instructions from the command and control module [6].

The acronym MOUT (Military Operations on Urban Territory) denotes military actions planned and conducted on a terrain complex where manmade constructions impact the tactical options available to commanders. Urban combat operations may be conducted in order to capitalize on the strategic or tactical advantages gained by the possession or control of a particular urban area or to deny these advantages from the enemy [4]. The characteristics of MOUT include complex situations brought about by engagements in urban environments (ambushes, civilians).

Combat Effectiveness (CE) can be defined as the ability of a (friendly) unit to rapidly and accurately sort and categorize detected objects (blue, white, red) and make a decision as to whether or not to employ deadly force against the identified object/target. Effectively applying the CE guarantees a minimum level of collateral damage and fratricide.

Now, to exemplify the previously defined terms, the following briefly examines Rules of Engagement (ROE). Together with tactics, techniques and procedures, ROE defines guidelines which then support an individual in a situation when a decision is made about whether or not to open fire. TTP supports the decision making process regarding force implementation in the Area of Operations (AOR). Depending on the ROE formulations, the orders concerning using force may vary as indicated in Figure 3.

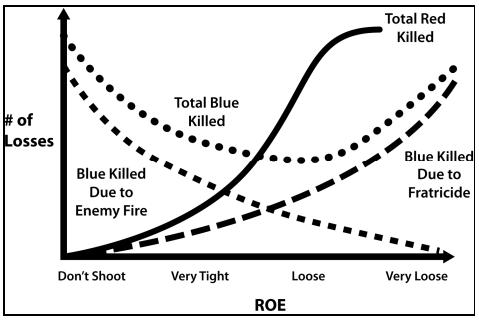


Figure 3. ROE in relation to the number of troops killed (blue, white, red) and the number of losses and fratricide [5].

All warriors depend on situational awareness (SA) [10] which can be provided also by using WPSN-systems introduced in [4]. The Blue Force Tracking-systems (BFT) along with the White Force Tracking (WFT) presented in [5] provide vital information for improving commanders' decision-making and avoiding fratricide and collateral damage. Blue Force (BF), allies and White Force (WF) need to be constantly precisely located.

II. RELATED WORK

This study has a strong linkage to dismounted FFW programs and Soldier Modernization Programmes (SMPs) ongoing in all major militaries. Obviously, they continue to be increasingly significant in enhancing the overall performance of militaries regardless of the financial retrenchment and downsizing demands. These programs concentrate on improving and updating dismounted soldiers' equipment thereby optimizing performance to minimize collateral damage and fratricide. The specifics related to the challenges concerning gathering and forwarding information are well known by militaries around the world. However, since nations invest significant sums of money on development projects, the specifics tend to remain classified and no valid test-data are available. The same applies to mathematical formulae, simulation results, and other types of ad hoc testing reports. Thus comparing and analyzing existing Future Force Warrior communication systems and architectures is currently a research mission impossible as the data remain unavailable for validation purposes. Developing military gear for future armed forces, and special operation forces in particular, continues to be expensive as devices are necessarily tailored for a limited number of users. This is why COTS-based gear solutions become an attempting alternative for military purposes as well.

Moreover, the existing COTS-based Command, Control, Communications, Computers, Intelligence, Information, Surveillance and Reconnaissance (C4I2SR) technologies with their applications are relevant in facilitating the developments necessary for overcoming the varying challenges encountered in the future battlespace.

In terms of practical battle proof examples, the technology for Blue Force Tracking (BFT) was used during Operation Iraqi Freedom (OIF) for coordinating operations among the Joint Services and with allies and resulted in reduced causalities due to enhanced SA [11]. Understandably, the means to increase SA via improved BFT and WFT are also developments in progress. Figure 4 illustrates an example of the dismounted FFW system from the perspective of selected warrior gear.

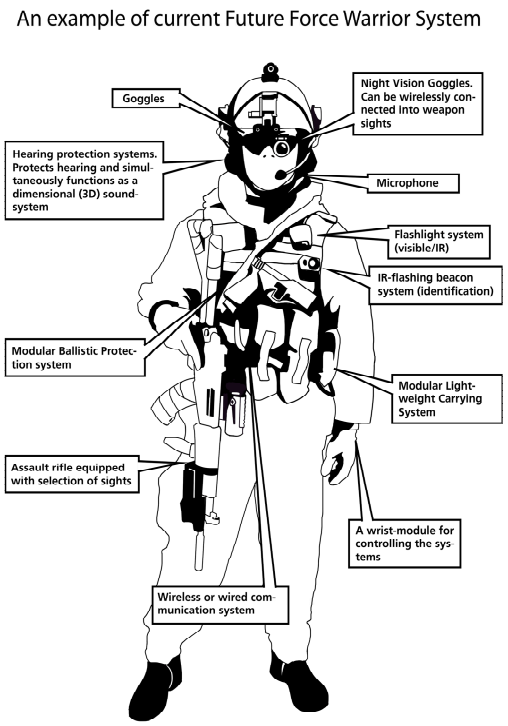


Figure 4. An example of a dismounted FFW with selected gear.

As introduced in [12], this paper continues to examine the interconnectedness of trained FFW and their gear in the light of the following three warrior levels: 1) the basic Warrior at the bottom level, 2) the Readiness Brigade Warrior, and 3) the Special Forces Warrior. The amount of TID and SA data varies along with the level of a FFW. In the higher echelons, the amount of data gathered via sensors and tracking systems is vast. To transmit and distribute the location information filtered and fused through various systems remains a challenge. A basic warrior located on the ground must fight rather than monitor his palm or wrist computer or lap-top.

In the building process of FFW, the key element is the hierarchy of the warrior levels (Basic, Readiness Brigade and Special Forces Warrior). As described in [12], concerning warrior levels, FFW act as moving relay stations to ensure the throughput of communication devices used. A warrior is in a key role in low level operations, and a warrior acts as a node or sensor in Network Centric Operations [12]. The FFW is an applicable sensor platform for Netcentric Operations as indicated in Figure 5.

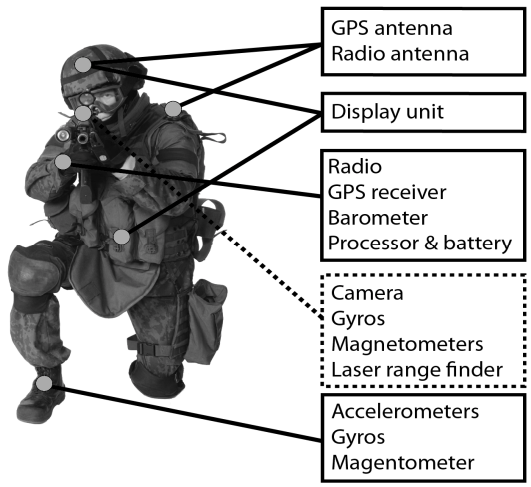


Figure 5. An example of a fully integrated Warrior for location purposes outdoors, indoors and in MOUT for contemporary warfare [19].

The principal contributing efforts, technical and procedural, involve the following [10][11]. First, CTI, automated query or response systems for dismounted personnel and light vehicles need to be addressed. Secondly, a means to share SA systems for employment at the platoon, squad, team, and individual levels must be applied. Thirdly, digitally-aided supporting fires' coordination and control must be defined. Fourth, Digitally-aided Close Air Support (DCAS) coordination and control has to be applied. In addition, challenges with Combat Identification Server (CIS) interoperability and Personnel Recovery Command and Control need to be solved. Lastly, marking and beacon systems for dismounted personnel, light vehicles, and friendly locations need to be applied. In fact, the US Army is fielding its new SA system known as Force XXI Battle Command and Brigade and Below (FBCB2) [10][12]. One of the keys into the success is careful mission analysis and thorough evaluation of Courses of Actions (COAs). Both processes can save time and minimize collateral damage. The use of available Blue and friendly Forces and resources can be optimized. This increases efficiency and along with minimum casualties, leads to minimum recovery times.

III. CHALLENGES IN COMBAT IDENTIFICATION

In military operations everything is done to prevent fratricide. Currently, identifying a warrior regardless of the visibility conditions is essential. As evident in Figure 2 earlier, both an interrogation unit and a responder unit are necessary, presupposing, first of all, that the systems are fully operational, and, secondly, that the distance between the warriors is appropriate. In case the identification system doesn't reply, a human is making the decision to open fire based on the TTP. The Identification to whether or not to open fire is based on the visual signature of the uniform, weapon and gear [5].

However, one needs to keep in mind that there is always the possibility that the location device gets stolen or misused by a third party in that, for example, an insurgent tries to function as a member of the White Force [5]. In order to increase the reliability of the system, the tracking devices have to be pre-coded and tied in pairs in advance before entering the battlespace to prevent the stealing of the tracking device. Once paired devices are torn apart, they stop functioning as planned – and devices become dysfunctional [5]. After the separation process, the devices must be re-paired and re-coded by the operator. During this process, the operator re-identifies the person.

IV. COMPREHENSIVE TARGETING PROCESS

First of all we start with the Point of Interest (POI) in the battlespace. When the POI has been detected, classified and finally identified, it may be indicated as a potential target. POI is not automatically a target. The utilization of Unmanned Vehicles (UVs) and Unmanned Aerial Vehicles (UAVs) and Unmanned Ground Vehicles (UGVs) play an important role in target acquisition, starting from the phase of detection of a POI.

When returning to ROE/TTP (whether or not to open fire), a link behind the targeting process deserves a closer look. The process is known as Detect, Identify, Decide, Engage, and Assess (DIDEA) [6]. The DIDEA provides an iterative, standardised and systematic approach supporting targeting and decision making, being generic enough to be used as a systematic process for Command and Control (C2) node targeting and decision making. Separate actions inside DIDEA area as follows [5]:

Detect: The process of acquiring and locating an object in the battlespace by analysing the phenomena in the electromagnetic spectrum.

Identify: The process of classifying an object into the category of blue, white (neutral) or enemy. This represents a primary step where specified CID tasks are accomplished.

Decide: The decision making process that follows the detection and identification phases. This is the most generic step within the process and represents the primary step where a specific ROE application occurs. In the decision-making phase, the executive officer / warrior has to decide and define what type of weaponry is appropriate for to the mission. In cases of opting for the use of deadly force, the following questions need to be addressed: 1. Can I engage (ROE application)? 2. If there are several targets, what is the order to engage the selected targets? 3. Which one is the most appropriate weapon system (most cost-effective, appropriate against the selected target)?

Engage: The execution of selected weapons in a selected order starting from the most dangerous target moving on according the panned sequence.

Assess: Monitoring the gained effects with the use of destruction power. Employing the force of various weapon systems available is repeatedly executed until the required level of destruction is achieved.

Once the critical data have been collected they have to be quickly analyzed to be used for evaluating different Courses of Actions. Success depends on an accurate mission analysis and a timely evaluation process of the accrued data. Improved SA results in optimal time for mission execution and simultaneous minimizing of casualties, which increases

efficiency and leads to minimum recovery times improving the overall efficiency of the troops utilized.

Once commanders have access to more current reconnaissance data for mission execution, they are able to analyze different COAs and, calculate the pros and cons to evaluate the best possible method to operate in any scenario prevailing. As explained in Figure 6, military commanders have by default value at least two different options for executing the mission in question. Once the Military Decision Making Process (MDMP) has been completed, the most effective operation can be executed to maximize the performance of the designated troops. In the described scenario below, the commander focuses the performance on destroying the Command Post (CP), the alternative number 2, instead of attacking against the armored enemy.

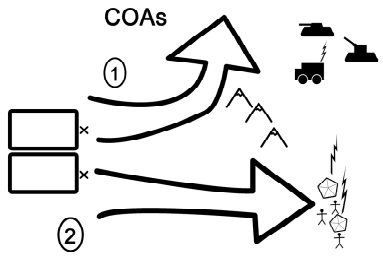


Figure 6. Possibilities of COAs [13].

When it comes to SSA, it is crucial to be able to distribute the accrued data rapidly and accurately in order to ensure success in military operations. When the accrued data remain intact and non-corrupted, both the execution of operations and the evaluation of COAs at all commander levels are improved. In particular in joint operations, the effective distribution of COAs and SSAs is in a central role.

V. How to Accrue Data for the Decision Making Process?

Self-evidently, cases of fratricide and collateral damage are bound to surface to some extent. Militaries are interested in locating both own troops and also increasingly the neutral entities of Non-Governmental Organizations (NGOs) and Governmental Organizations (GOs), the WF, the members of which can be tracked by using WFT described in [1].

Briefly put, the problem relies in relating the TTP, CID, COP and SA to the rules of ROE. This involves dealing with the balance described in Figure 3 earlier. If ROE formulations are too strict, for example, the commander's intent is to avoid the use of deadly force unless it is absolutely certain that the targeted object is positively identified to be an enemy – the Blue Force will suffer on the basis of the actions caused by the enemy. And, if ROE formulations leave too much room for interpretation, various types of casualties (red, blue and white) are bound to occur. Thereby the transmission of combat-critical location and identification data plays a crucial role in the battlespace.

The process of a complete targeting process can be described in a simplified form in a formula: Detect, Identify, Decide; Engage and Assess [6]. The DIDEA provides an iterative, standardised and systematic approach supporting targeting and decision making, being generic enough to be used as a systematic process for C2 node targeting and decision making. This process is thoroughly discussed in [6].

Furthermore, older existing systems are available for distributing data gathered by various types of sensors in various types of military and humanitarian crises environments. These technologies are based on WPSNs described in [4] and Wireless Sensor Networks (WSNs) described in [14][15]. The former are passive and will remain hidden whereas the latter are active and represent a more easily detectable system. Both systems are applicable in transmitting constantly flowing data from a sensor to a node, for example, to a vehicle or an unmanned vehicle (UV).

As suggested in [12], viable COTS-based methods exist, which improve the C4I2SR of a warrior at all the levels. The examples covered are based on usability cases of WPSN-solutions. They indicate that a warrior can obtain more critical information on the battlespace by using the presented WPSN solutions. This improves the general efficiency of a warrior at all levels. The platforms used today on the battlespace are not efficient. This is because they are based on a single sensor and they do not collect data in a way that would allow collaboration of multiple sensors. The proposed solution makes use of multi-sensor collaboration for improved location information and improved SA. Figure 7 explains the structure of a warrior skeleton as well as the location of the WPSN-system inside the FFW-system [4].

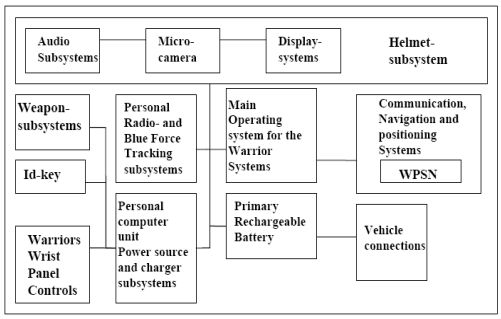


Figure 7. A Warrior's electronic skeleton [4].

In terms of FFW equipment, warriors need to be functional and their gear must be planned according to the set tasks. A key factor is the efficiency of a warrior, which can be gained via an improved SA, BFT and Command and Control. Warriors have to maintain their agility and remain active in the battlespace. However, since only combat-crucial gear can be hauled along, thereby not nearly all the gear necessary can be attached to the dismounted FFW. Thus the warrior skeleton and its communication systems need to be carefully defined and built at each warrior level according to the given task requirements. Currently, the present solutions seen in active use are cumbersome and lack integration. The WPSN-solutions still remain unapplied in these platforms. Thus the maximum potential remains unreachable without effective sensor and data fusion. Militaries are moving towards smaller specialized units while the overall performance requirements keep increasing. At the same time

troops are designed and trained for dismounted operations in which a greater degree of flexibility and reliability of battle-proof and robust systems are needed.

Practically speaking, small militaries are often unable to utilize the possibilities in target acquisition offered by UVs and UGVs. As introduced in [13], affordable and easily deployable Sensor Munition Element (SEM) offers new possibilities to accrue data behind the enemy lines. The system is based on Commercial-off-the-shelf (COTS)-products and is affordable for the use of small troops utilized in small militaries.

Means to accrue SA-data in the battlespace are depicted in Figures 8 and 9. All available means are utilized in order to avoid fratricide and collateral damage thus maximizing the performance of own troops to ensure mission success.

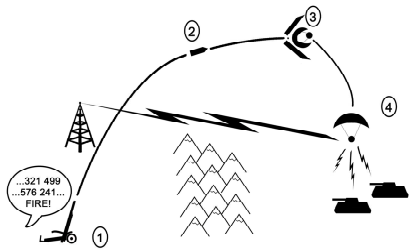


Figure 8. On deploying an SE above an enemy territory: 1) Fire Support Order is commanded, 2) SEM is airborne, 3) SEM opens and ejects the SE, 4) the SE starts to transmit gathered data from the enemy territory and targets [13].

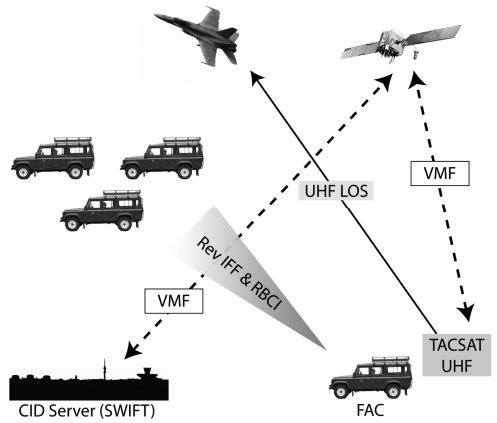


Figure 9. The Comprehensive system of gaining SA-data to avoid collateral damage and fratricide [5].

The decision as to whether or not to open fire is based on the visual signature of a given uniform, weapon and gear as well as magnetic, seismic or acoustic signals identified by a sensor [14] as described earlier. Self-evidently, the transmission of combat-critical location and identification data play a crucial role in the battlespace. Once the accrued data have been transmitted and received, they flow through a dissemination process, where these data are analyzed and fused to form a COP and to increase the overall SA. Figure 10 explains the process of Signature Prediction Process (SPP).

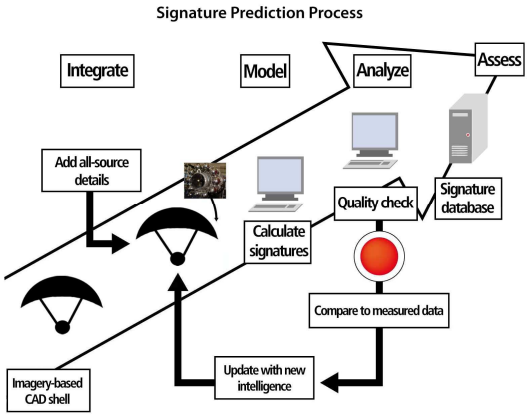


Figure 10. The Signature Prediction Process, typical of several available surveillance and detection systems.

The destruction power of a given weapon system has to be optimized according to the enemy location (forest, open area, Urban Territory), the state of movement on-the Move (OTM) or at-the-halt (ATH), and the protection-level (mounted, dismounted, dug). Apart from this, the commanding officer must keep in mind that operations are executed with improper SA, COP and suffer from lack of precise real- time CID.

Figure 11 emphasizes the importance of SA around the target area. The shooter has to be aware of the locations and status of both own troops and the enemy. It is critical to optimize the destruction power of a weapon system along the identification of a target. When the target represents a hierarchically critical enemy commander, he or she can be destroyed by transmitting the coordinates and visual signature to the designated shooter, as indicated in Figure 11.

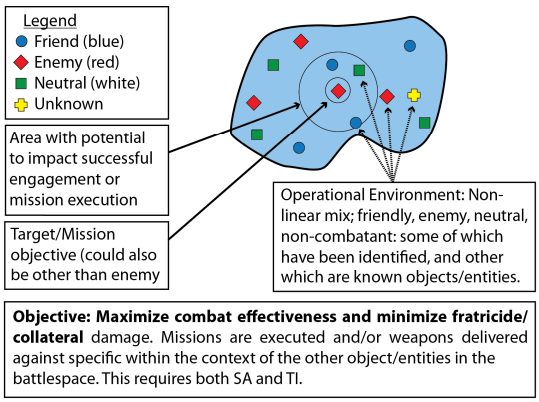


Figure 11. The importance of the SA around the target area [5].

To enhance improved SA and COP, Geographical Based Situational Awareness (GBSA) can be utilized [14]. The system utilizes the VHF-frequency operated Combat Net Radios (CNRs). When the CNRs are on the connectivity range, they recognize and identify radios in the system. Once the radios are at the same channel and the clock (hopping sequence) of CNRs are in a correct time, a reliable SA-tool

[14]. The main problems related to this system have to do with the clock and hopping sequence. This is one possibility to minimize fratricide and collateral damage. At the moment, the main benefit of this concept is in preventing from being fired at by own weapon systems, minimizing incidents of fratricide by means of improved SA-information.

VI. Challenges Involved in Distributing Situational Awareness Data

The amount of data accrued via versatile sensors and tracking systems is necessarily immense to say the least. As a result, to distribute the location information filtered and fused through various systems remains a challenge. As said, warriors' main function remains to fight instead of double-checking monitor his palm-top or equivalent. Besides, there will always be disturbances in electromagnetic spectrum, quality of service (QoS) and transmitting power along with the limited bandwidth set limitations to the ubiquitous communication systems. As indicated in Figure 12, the possibilities of battlespace communication are versatile, since almost all the sensors utilized are somehow linked together to facilitate BFT and CID and to improve COP and SA.

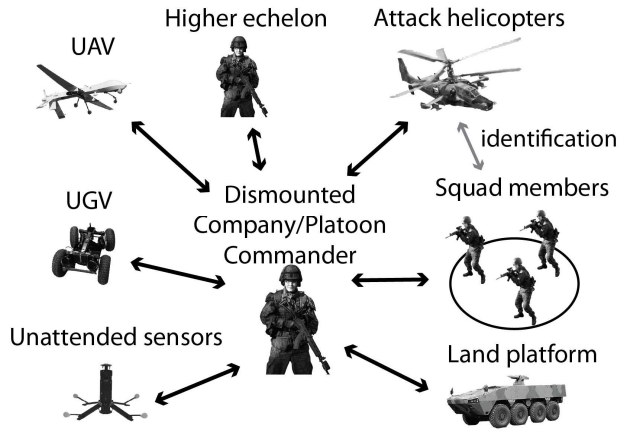


Figure 12. The types of possible platforms serving as sensors and network nodes [5].

The problems encountered in data distribution are linked to the present existence of various devices and data in interfaces. BML can be seen as a common language enabler between gadgets and interfaces [2] along with almost ubiquitous swarms of UAVs described in [15]. Limitations in energy and bandwidth play a vital role. The locating of instruments of various types consumes reasonable amounts of energy, not to mention the increase in weight and number of devices in warrior gear and required maintenance. Due to lack of accessible wireline infrastructures, unmanned systems have to be powered through a combination of batteries, solar power, and power scavenging [16]. When FSO-technology is adopted in backbone networks and between selected ground stations, an intelligent, dynamic and secure data transmission with high data rates can be offered to mobile end-user [17]. FSO-technology offers high-speed, reliable and cost-effective connectivity for heterogeneous wireless services provision in both urban and rural deployments when Dense Wavelength Division Multiplexing (DWMD) is utilized in Radio-on-FSO (RoFSO) system [18]. It has been demonstrated in tests that the advanced DWDM RoFSO offers a viable solution to provide broadband wireless connectivity. Radio over Fiber (RoF) technology will most likely offer a reliable data transmission rate of 10 Gbps in the next generation FSO-systems [18]. A simplified principle of FSO-communication system is introduced in Figure 13.

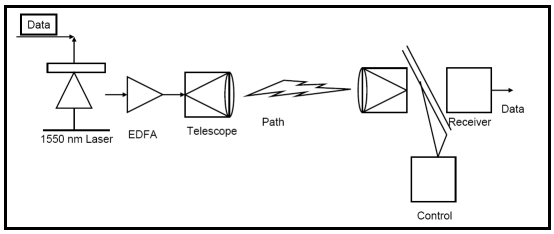


Figure 13. A schematic diagram of a point-to-point FSO communication system [19].

To maximize the possibility of devices communicating in a proper and planned manner, the topology of network systems has to be correctly coordinated (manage spectrum usage with group mobility patterns) [20]. In addition, the hierarchy of a network has to support and enable this. Both the goals can be achieved by hierarchical design where devices are only to interact with their peers from the same group [21]. Furthermore, the transmit antenna selection is a practical technique for achieving significant power gain, even with commodity hardware and without changes to different waveform protocols [13].

As discussed in light of usability cases presented in [16], WPSN is beneficial because of the following reasons: the effect of roadside bombs can be avoided once their precise location is known early and precisely enough. The increased knowledge at the basic warrior level in the form of location information gained from the Self-Calibrating Pseudolite Array (SCPA) on the battlespace improves warriors' ability to carry out the set tasks. Roadside bombs can be detected early enough and dismantled or destroyed before own or allied forces arrive on the spot. The Special Forces utilize the same output of SCPA while conducting their ultimate tasks. Since the nodes of WPSN do not communicate with each other, the system remains concealed, yet active. The WPSN node communicates with a UAV through encrypted messages. Thus WPSN responds only after a UAV has submitted a polling request with a specific code. Utilizing swarms of UAVs and UGVs has to be emphasized. The routes of UVs can be fed into the systems early enough to gain the needed information from the designated areas as depicted earlier in Figure 11.

VII. Location and Communication Possibilities in Urban Areas

An Army tactical warfighter needs network services both OTM and ATH [5]. One of the lessons learned from Iraq and

Afghanistan was the need for a more robust Beyond-Line-Of-Sight (BLOS) communication capacity between the lower Army echelon Land Warriors, from Squad Leaders to Battalion Commanders [5].

The proposed and described solutions have to be based on novel, generic and robust battlespace-proven solutions in order to meet the given needs, and this in turn involves addressing the topology of the network system carefully. In MOUT transmitting and receiving signals of different waveforms simultaneously is challenging due to the nature of the combat environment [22].

Since the power production and power consumption will remain as a challenge, certain issues need to be addressed. Thus when defining the network design, it has to be emphasized that network coding enables a more efficient, scalable and reliable wireless network [23].

The MOUT environment features no service of the Global Navigation Sensor System (GNSS) indoors, and indoors propagation poses a serious problem. The placement of an antenna platform is challenging. One solution can be the installing of a high-bandwidth conformal antenna in the soldier's helmet with the coverage of over 750 MHz through a 2,7 GHz frequency band [24]. The combat-critical solutions involve improving communicating, SA and transmitting C2 information among highly dispersed battlespace units in dynamic environments, such as MOUT [23] [24].

Next, let us assume that there is a WPSN-system available for positioning and location services. If the capability of GPS-Pseudolite, better known as the Self-Calibrating Pseudolite Array, is attached into the satellite-based Carrier-phase Differential GPS-type (CDGPS), it is possible to determine positioning in locations without access to the GPS satellite constellation [4] [15] [25]. This will improve locating own troops inside buildings dramatically, thereby significantly improving CID, TID and SA. This system is depicted in Figure 14.

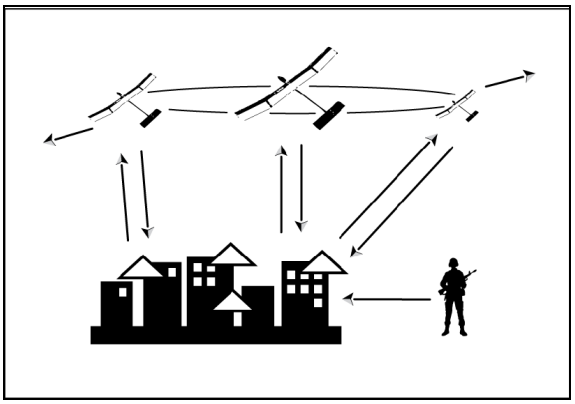


Figure 14. The WPSN presented in the urban infrastructure [4].

VIII. DISCUSSION

This study examines COTS-based communication technologies available for increasing dismounted FFW performance, minimizing collateral damage, improving SA and COP and focuses on how to apply technical solutions in the military environment to enhance the overall performance of a dismounted FFW. Applying suitable and relevant applications, C4I2SR tools in the existing networks, aids in overcoming the varying challenges in the battlespace [4][26]. Possibilities of Free Space Optics (FSO) can be utilized, as introduced in [19]. This study outlines aspects of applying the existing communication technologies, C4I2SR, to military battlespace systems [3]. In addition, the C4I2SR tools for dismounted FFWs have to cater for the requirements of affordability, reliability, versatility and modularity [4].

Means to present the accrued data are versatile and challenging. Since, as can be repeated ad nauseam, a warrior must primarily fight, the chosen method to present fused data has to support warriors' main task rather than disturb and distract. Especially, in order to be able to present SSA-data appropriately in battlespace settings, the assisting role and practical features of Graphic User Interfaces (GUIs) practically remain utterly important. This practical usability angle is depicted in Figure 15.

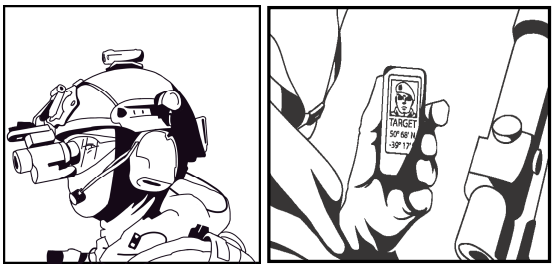


Figure 15. Means to forward the accrued data via various warrior displays.

In terms of the equipment angle to the FFW-concept, a properly equipped FFW represents a warrior who is supplied with the latest technology applicable which translates into enhanced performance capabilities in versatile terrain, including MOUT, CME and special operations. As evident, this asks for computer-aided modularity and scalability to allow for adaptability according to warriors' task-levels, timings, and locations of operations. Furthermore, the integration of subsystems must be possible in order to ensure the optimal functionality and accurate data transmission between the given systems. This requires that the equipment be rapidly replaceable and exchangeable for the purposes of location services and C4I2SR -systems.

As denoted in [7], asymmetric warfare sets more challenges compared to traditional warfare. This involves the challenges related to identifying the Point of Interest (POI) in the battlespace. It is essential to define the POI early enough as an enemy (red), own (blue), or neutral, White Force [27]. Before the execution of weapon systems, the commanding officer and a single Warfighter has to be in control of the given situation to avoid fratricide [28]. In case the POI is identified as an enemy, the decision of possible use of force has to be made rapidly [29].

An identification device utilizable in a battlespace consists of a transmitter and the receiver elements, the

former based on laser, the latter on a radio frequency (RF) system. Warriors can be equipped with Cooperative Target Identification Systems. CTI allows a human shooter or sensor to interrogate a potential target and thereby forces the potential target to respond to the interrogation in a timely manner as described earlier in Figure 2 [5] [7].

As concluded in [12], the equipping of a FFW can be pictured by means of a product line warrior drawing from three-tier warrior levels. An FFW's gear has to be designed to meet the requirements set by the future hybrid battlespace [30]. Therefore, the warrior equipment must be versatile and modular. Moreover, remotely controlled UVs serve as tools to improve SA and BFT, and thereby assist in ensuring mission success [31]. The number and nature of different Human Machine Interfaces (HMIs) is growing with an increasing speed.

A computer can be adopted in varying roles depending on the warrior level in question: the computer can be mounted to clothing or on the wrist, for example. The higher the role of a warrior, the more a computer is seen as an assistant. In contrast, the lower the level, the more the computer forwards tasks. As presented in Figure 16, a computer can be programmed to task a warrior to move and fight at a certain pace depending on the mission. A computer can command a warrior to move at a certain pace and directions following the cycle of friendly fire missions as indicated in Figure 16. This process increases warrior efficiency, minimizes fratricide and increases a commander's SA.

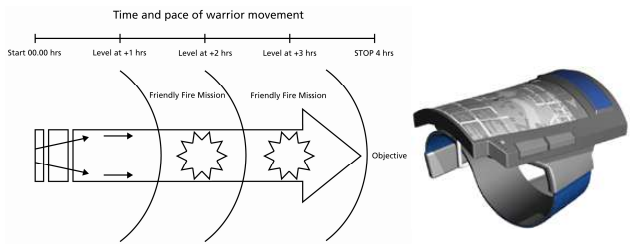


Figure 16. The principle of computer-tasked pace of movement (r) and a wrist module (l) [4].

Effect-based thinking and systems engineering serve as the tools to be deployed to achieve the ultimate goal: the optimally functioning effective FFW at all the command levels in all potential battlespace environments. Remotely controlled UAVs and UGVs can act as assisting tools for a warrior [32]. They can facilitate BFT and improve SA thereby increasing the probability of success in missions, even when operating Beyond Line of Sight (BLOS) [31].

According to [4], warriors have to remain functional and their gear needs to be planned according to their set tasks. A key factor is the efficiency of a warrior, which can be gained via improved SA, BFT and C4I2SR [12]. A Warrior has to maintain his or her agility and remain active in the battlespace. Furthermore, only part of the gear necessary can be attached. As demonstrated via the usability cases, WPSN solutions together with SCPAs and UVs can be utilized to reach the maximum performance at all the warrior levels. Planning the warrior's gear requires a deep understanding of the operational environments and the requirements set for the performance of a warrior.

As introduced in [12], the issue of exploiting C4I2SR-tools is not only a matter of a complicated command and control system [33]; it is a matter of trust in the entity, especially in operations utilizing collaboration tools of various types [34]. Each entity embedded into the C4I2SR-tool environment can contribute added value to SA and intensify the desired outcome by committing themselves to and abiding by the set rules and policies. Only this way C4I2SR -tools can be maximally exploited, and increase the number of promising instruments for the enhanced performance in ME and CME.

By being successful in merging all these described elements and tools, several C4I2SR -related challenges can be solved. As long as a human being serves in a loop as a performing entity, there will always be a certain amount of mistrust. Finally, once a reliable tool for distributing traceable tasks can be created, the amount of trust between entities can be increased.

As in [35], all the entities need collaboration for their mission success and survivability in ME and CME operations [36]. If an entity fails to collaborate, it takes a calculated risk to fail. Collaboration requires suitable tools and reliable and ubiquitous network systems [37]. Collaboration is necessary for avoiding chaos and avoid wasting resources in order to combine resources for an optimized outcome [37].

As demonstrated in [35], three results are offered as a contribution for the further development of Command and Control-tools: 1) a C2-tool, which enables use of Business Process (BP) in the command and control process; 2) the Resource Manager (RM), which is a central element of the Military Service Oriented Architecture (MSOA) in the distributing of limited resources; lastly, 3) the BP in the ME along with the MSOA [35], [38]. These results offer the yet missing attributes for the C2-tools for ME and CME. Combining these elements enables a successful control for the BP in ME and CME settings. Furthermore, [35] introduces the composition of the RM and the role of a scheduler, the function of the BP, and highlights the significance of trust and commitment in CME [35]. Trust is needed to gather information of the entities and to ensure tasks will be completed in the given time and manner [35]. Each entity embedded into the C2-tool environment can add increased value into the SA and thereby intensify the outcome by committing themselves to rules and abiding by the set policies. Understanding the meaning of combining the presented new tools gives an edge in the battlespace to perform more efficiently and with a minimum number of casualties.

As denoted in [19], FSO-technology offers a secure and reliable means to forward a constant flow of data with an adequate transmission rate [17]. Present communication systems on a warrior level are energy consuming and require a lot of training in order to benefit from the system [39]. The FSO-system in turn is simple to use, and thereby also less trained FFWs can effortlessly perform the necessary communication tasks. The overall reconnaissance system

benefits from FFWs, individual sensors, sensor networks, and mathematical analyzing and data mining programs, resulting in high level data for increased SA [40]. The key function of an FFW is to collect large amounts of SA information and forward these data to the CP for further data analyzing processes. In brief, adopting FSO in active use allows for a system featuring high transmission security, high bit rates, low bit error rates, and no need for expensive optical or copper cables [41] and FSO can be utilized when using DWDM as introduced in [18]. The main limitations of FSO-technology are related to its susceptibility to the effects of atmospheric absorption, smoke, rain, fog, snow (attenuation), and pollution/smog and, obviously, a free line-of-sight [42]. These factors restrict FSO devices' range communication capability to cover approximately a mile in optimal conditions [41].

TABLE I. THE SYSTEM CHARACTERISTICS OF AND FUNCTIONAL REQUIREMENTS FOR THE FSO COMMUNICATION SOLUTION [19].

System Characteristics	Functional Requirements
Communication	High bit-rate, urban range in the usage scenario, hard to intercept and detect
Physical	Lightweight, low energy consumption, quick set-up
Architectural	Modularity, versatility, based on existing gear, suitable to Network Enabled Defence (NED), IP interface
Economic	Affordable, disposable, COTS-based
Dependability	Reliable, secure, proven technology
Capability improvement	Addresses a realistic capability gap (many relevant scenarios)

The WPSN-solution features many advantages over those of the traditional WSNs. This is, polling can use sensor specific codes and thereby security issues become easier to tackle [4]. Moreover, the energy consumption of the nodes in the fixed network is more equal since multi-hop data transmission is removed. The fixed sensor nodes do not lose connectivity even if a large number of nodes is removed [4].

As demonstrated via the presented usability cases [4], WPSN solutions together with SCPAs and UVs can be exploited to reach the maximum performance at all warrior levels. Planning an FFW's gear requires a deep understanding of the environment and the demands set for a warrior. The warriors' niche and the nature of their missions have to be thoroughly understood. The keys to success rely on precise planning based on the needs of warrior systems and subsystems from bottom to top

Obviously, in all military operations and especially in low-level tactical military operations in particular, critical Situational Awareness data have to be collected rapidly, since mission success is time-dependent. Figure 5 concentrates on describing the data accruing process, when Figure 6 expresses the outcome of MDPM as alternative COAs. Once data have been accrued, a battle can be won only by careful mission planning, comparing different COAs and rapidly executing successful operations.

Figure 8 earlier illustrates the process of targeting by utilizing the capabilities offered by the COTS-based SEM. Thereby the adoption of existing COTS-technologies and their solutions, when appropriately applied, offers a key to ensuring the desired success.

Since accurate and timely identification in the battlespace is a matter of life and death for each warrior, a careful analysis of the performance and capabilities of chosen systems needs to be carried out before introducing these systems in the battlespace. CID equals the process that warriors and sensors go through in order to identify battlespace objects prior to deciding whether or not to open fire. Warfighters are trained to employ all available means at their disposal to define and assess potential targets in the battlespace prior to applying combat power. CID can be seen as a complex series of networked systems, procedures and doctrine as presented in Figures 1 – 6. These systems also include the definitions of TTP, COP, SA, ROE and DIDEA.

More specifically, problems can arise in particular in commanding and being commanded. A Combat Identification Server (CIDS) offers military commanders and warriors access to accurate and near real-time BFT and WFT systems [5]. Besides this, CIDS offers commanders a tool which can foster improved mission planning resulting in increased accuracy and tempo of missions. To sum up, CIDS aids commanders to reduce the number of unexpected incidents and minimize collateral damage.

Once the TTP, CID, COP, and SA systems discussed in this paper (cf. Figure 1) are designed, tested and become fully implemented as part of the combat gear, some progress may be discernible in minimizing fratricide and collateral damage. The reality is that for as long as human actors remain part of any decision-making processes, incidents of fratricide and collateral damage are bound to occur. All efforts to minimize the human error factor by improving existing technologies, TTP, CID, COP and SA together with defining explicitly the formulations in ROE, are to be saluted. The efforts to minimize the unwanted phenomena are to be applied, for instance, in an ongoing series of Bold Quest exercises.

IX. CONCLUSIONS

In the very beginning of this study, the following three questions were raised: 1) How to optimize the performance of a dismounted FFW by means of improved SA? 2) How to increase SA with the available COTS-based communication technologies? 3) What are the means to avoid casualties, collateral damage, and fratricide?

First, as for the question of optimized dismounted FFW performance, regardless of the asymmetric and hybrid characteristics of future wars and conflicts with their respective battlespaces, combat settings necessarily involve a duel: participants try to surprise and outwin each other in terms of positions, timings, maneuvers, and technical capabilities. Defining suitable technological solutions as part of the FFW gear ensures the optimal FFW performance, which presupposes reliable and technologically mature C4I2SR-tools suitable for use in various battlespace environments with ubiquitous communication data transmitted with the solutions of NCW. An FFW functions in NCW contexts as a force multiplier of the network centric

C2-cycle from the sensor to the shooter aiming at minimized numbers of fratricide and collateral damage.

As for the elements necessarily part of the FFW-concept, the following features become seminal. FFWs with their computer-aided equipment need to be designed to meet the battlespace requirements dependent on their respective niche. Thereby the designing and constructing process of the FFW must be taken into account. To simplify, each warrior acts as a node or sensor and thus needs a reliable, versatile, modular and scalable electrical platform to receive and transmit the necessary data and information in a given operation- and task-dependent timeframe. Moreover, it needs to be possible to integrate subsystems to ensure the optimal functionality and accurate data transmission between the given systems, which in turn enhances overall warrior performance capabilities.

FFWs' functionality aims at improved overall performance and situational awareness (SA), which become evident in, for example, the instances of Blue Force Tracking, and Combat Identification (CID) facilitated by the capability to utilize data transmitted by UAVs and UGVs.

FFWs' personal computers' status is again dependent on the warrior-level in question: slave, assistant, or master. The role of the computer in all warrior levels is to enhance the overall SA, avoid fratricide and collateral damage and lastly but least, improve the performance of the warrior. Once the system that still currently remains to be designed is fully operational, a computer may order warriors to carry out an offensive in a particular direction at a given point in order to maximally utilize their performance capabilities. Since terrain requirements vary from remote locations or densely built-up areas to versatile battlespace, all the warrior equipment must be adaptable and able to support the warrior in the changing circumstances.

Obviously, the overall objective of planning and designing an optimally functioning FFW aims at avoiding fratricide and minimizing collateral damage. This ensures that all the resources available are focused on getting the ordered tasks fulfilled maximally. The end result then equals a state of Combat Effectiveness that enables a given unit to rapidly and accurately sort and characterize detected objects into relevant categories (blue, white, red), and, consequently, make a decision as to whether or not to employ force against the identified object / target.

Second, when it comes to improving SA by means of utilizing COTS-based communication technologies, challenges in CID continue to surface. As discussed, solutions for pinpointing and locating POIs can be based on COTS-technology. Yet, although the required technologies do exist, their usability still has to be tested and re-evaluated, and thoroughly selected solutions need to be adopted to avoid unnecessary casualties and destruction in the battlespace.

In terms of targeting, a more comprehensive targeting process can be attained with the assistance of UVs and UAVs. Once the targeting process is effective, it boosts the DIDEA decision making cycle. Moreover, by improving the targeting process, tools for better decision making can be offered. This in turn results in a better analyzing of COAs. Applying and properly executing the most favourable COAs facilitates are mission success with minimized number of fratricide incidents and collateral damage.

New technologies, such as FSO combined with WPSN can improve TID, CID, COP and SA. FSO offers a possible means to transmit large amounts of data to Command Posts with quick wireless set up. FSO also offers an insensitive and reliable means to improve the overall SA in ME and in CME. These all together support mission success and improves the overall efficiency in execution of versatile operations in rapidly changing operational environments.

As for the overall FFW gear development, extensive field trials with actual troops are required in order to test, validate, and evaluate the performance of the C4I2SR-gear. The focus of using these tools has to be in detection, identification and target acquisition processes. The capability to embed the required gear on the warfighter has to be combat proof. The interfaces between the human and the machine have to be designed, tested and evaluated to determine the optimal solutions to meet the set objectives. To ensure data distribution between various platforms, interfaces and machines problems in data distribution are linked to various devices. As noted earlier, BML can serve a common language enabler between machines and interfaces as well as a tool in exchanging data between and among swarms of UAVs.

In the future, the overall troop performance aided by assisting electrical devices has to be evaluated in varying environments, such as open terrain, MOUT, desert and forest terrain, and multiple scenarios have to be exploited as test-beds for realizing improved COP and SA. The level of the adopted gear has to match the existing chain of command and the task-based level and capability of the performing troops. Furthermore, the development of the user interface for the UVs remains a challenge. In addition, Graphical User Interfaces (GUI) are significant in maximizing the potential of the adopted and implemented gear. Thus extensive series of tests both in laboratories and as field trials are required to optimize the user-friendly GUIs.

There is an ever increasing need for more effective and versatile warriors. Armies of the world are downsizing their number of troops while requiring increased performance of the remaining military power, and, ever increasingly, versatile tasks along warfighting, including executing humanitarian missions, continue to set new requirements for warfighters and their capabilities.

And, finally, the bottom line here obviously targets the question of how to minimize casualties, collateral damage, and fratricide. As CID and TID systems continue to remain inadequate for battlespace settings, new COTS-based technologies and applicable solutions are both welcome and indispensable. As we speak, all the decisions as to whether or not to apply combat power boil down to a human being executing the decisions and owning the ensuing actions. Therefore any affordable means available must be exploited in order to be able to resort to applications and gear which truly facilitate improving the performance of dismounted FFWs, optimizing SSA and thereby reducing the number of

instances which inevitably feature lives and assets lost no matter how honed the gear and minds involved.

REFERENCES

[1] T. Saarelainen, Enhancing Situational Awareness by Means of Combat-ID to Minimize Fratricide and Collateral Damage in the Theater, The Sixth International Conference on Digital Telecommunications, ICDT2011, April 17 – 22, 2011 Budapest, pp 131 – 139.

[2] K. Rein, U. Schade and M. Hieb, Battle Management language (BML) as an enabler, IEEE Conference on Communications, ICC'09, 14 – 18 June, Wachtberg, Germany, 2009, pp. 1 – 5.

[3] R. Buchanan, Intelligence as Fuel Data Requirements for Modern CID Systems, A presentation in Combat-ID and Situational Awareness Conference, IQPC, London 16.11.2010.

[4] T. Saarelainen, and J. Jormakka, C^4I^2-Tools for the Future Battlefield Warriors, in Proceedings of *IEEE International Conference on Digital Communications (ICDT)*, 13 –19 June 2009, Athens, Greece, pp. 38 – 43, doi 10.1109/ICDT.2010.15.

[5] T. Saarelainen, Enhancing Situational Awareness by Means of Combat-ID to Minimize Fratricide and Collateral Damage in the Theater, The Sixth International Conference on Digital Telecommunications (ICDT2011), April 17-22, Budapest, Hungary.

[6] E. Ospital, A. Wojack, A Holistic Approach to Combat Identification, Military Technology, vol XXXIV, issue 9, 2010, ISSN-3226, pp. 68 –73.

[7] T. Saarelainen, White Force Tracking, in Proceedings of the *8th European Conference on Information Warfare and Security*, *ECIW2009*, 6 – 7 July 2009, Lisbon, Portugal, pp. 216 – 223.

[8] J. Donner, Combat Identification with German Air Force, A presentation in the Combat-ID and Situational Awareness Conference, IQPC, London 16.11.2010.

[9] Bhattacharyya, A., Mazumdar, C., and Saraswat, V., A Real-Time Robust Decision Support System for Network Centric Warfare, in *Proceedings of the 8th European Conference on Information Warfare and Security*, *ECIW2009*, Braga, Portugal, 6-7 July 2009, pp. 9 – 19.

[10] K. Chevli, P.Kim, A.Kagel, D. Moy, R. Pattay, R. Nichols, A. Goldfinger, Blue Force Tracking Network Modeling and Simulation, IEEE Military Communications Conference, MILCOM 2006, 23-25 Oct. 2006, Washington, U.S.A., pp. 1 – 7.

[11] Bold Quest 2011 Overview, a presentation attached to the Finnish Defence Forces' Letter of Intent concerning the Bold Quest 2011.

[12] T. Saarelainen, and J. Jormakka, (2009) *Computer-aided Warriors for Future Battlefields*. Lisbon, Portugal: ECIW2009 Press, pp.224–233.

[13] T. Saarelainen, Targeting Situational Awareness beyond the Event Horizon by Means of Sensor Element Munition, submitted for IARIA2012 Conference.

[14] T. Sierksma, J. Hoekstra, B. Jansen, Geographical Based Situational Awareness in Military Mobile Domain, IEEE Military Communications Conference, MILCOM2007, 29 – 30 Oct, 2007, Orlando, U.S.A, pp. 1 – 7.

[15] D. Al-Abri, J. McNair, Improving Localization Accuracy in Wireless Sensor Networks using Location Verification Feedback, MILCOM 2007, 29 – 31, Oct. 2007,Orlando, FL, USA, pp. 1 – 7

[16] O. Kosut, A. Turovsky, J. Sun, M. Ezovski, L. Tong, and G. Whipps, Integrated Mobile and Static Sensing for Target Tracking, MILCOM 2007, 29 – 31 Oct. 2007, Orlando, FL, USA, pp. 1 – 7.

[17] A. Desai, S. Milner, Autonomous Reconfiguration in Free-Space Optical Sensor Networks, *IEEE Journal on Selected Areas in Communication, vol 23, issue 8, 2005*, pp. 1556 – 1563.

[18] K. Kazaura, K. Wakamori, M. Matsumoto, N. Higashino, K. Tsukamoto, and S. Komaki, A Proposal for a Broadband Wireless Access Technology based on Radio-on FSO Links, in Proceedings of *IEEE GLOBECOM Workshops*, Nov. 30 – Dec 4, 2008, New Orleans, U.S.A., pp. 1 – 6, doi 10.1109/GLOCOMW.2008.ECP.71.

[19] T. Saarelainen, Free Space Optics in the use of the Future Force Warriors, International Society in Military Studies (ISMS), 10 – 11.11.2010, Stockholm, peer reviewed, accepted and presented. This article is to be published in an issue of the Journal of Communications and Computer in 2011**.**

[20] T. Saarelainen, and J. Jormakka, Computer-Aided Warrior for Future Battlespaces, Proceedings of the 8th European Conference on Information Warfare and Security, ECIW2009, Braga, Portugal, 6-7 July 2009, pp. 224 – 233.

[21] H. Zheng, J. Shi, L. Cao, Group-Mobility-Aware Spectrum Management for Future Digital Battlespaces, MILCOM 2006, Washington, DC.

[22] J. Miller, Coalition Combat Identification Capability Assessments "Bold Quest", A presentation in the Combat-ID and Situational Awareness Conference, IQPC, London 16.11.2010.

[23] G. Jakobson, J. Buford and L. Lewis, A Framework of cognitive situation modelling and recognition, MILCOM 2006, 23 – 26 Oct. 2006, Washington, DC, pp. 1 – 7.

[24] D. Herold, L. Griffiths, and T.Y. Fung, Lightweight, High-Bandwidth Conformal Antenna System for Ballistic Helmets, MILCOM 2007, Orlando, FL, USA.

[25] M. Matsuoka, S.M. Rock, and M. G. Bualat, Autonomous Deployment of a Self-Calibrating Pseudolite Array for Mars Rover Navigation, Position Location and Navigation Symposium, PLANS 2004, 26 – 29 April 2004, pp. 733 – 739.

[26] K. Lund, A. Eggen, D. Hadzic,T. Hafsoe, and F.T Johnsen, Using web services to realize service oriented architecture in military communication networks, *Communications Magazine, vol 45, Issue 10*, pp. 47 – 53, doi 10.1109/MCOM.2007.4342822.

[27] H. Chang, and J. Zhang, New metrics for clutter affecting human target acquisition, *IEEE Transactions on Aerospace and Electronic Systems, vol 42, issue 1*, Jan. 2006, pp. 361 – 368, doi 10.1109/TAES.2006.1603429.

[28] W-h. Xu, Y. Zhang, and Z-X. Li, Lightweight weapon system for robot soldier, in Proceedings of *4th IEEE Conference on Industrial Electronics and Applications, 2009, (ICIEA 2009)*, pp. 1727 – 1729, doi 10.1109/ICIEA.2009.5138492.

[29] T.W. Hayden, and C. Ward, Human collaboration tools for net-centric operations, in *Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems*, 20-20 May 2005, St Louis, MO, pp. 20, doi 10.1109/ISCST.2005.1553286.

[30] N. Suri, E. Benvegnu, M. Tortonesi, C. Stefanelli, J. Kovach, and J. Hanna, Communications middleware for tactical environments: Observations, experiences, and lessons learned, *Journal of IEEE Communications Magazine, October 2009, vol 47, Issue 10*, pp. 56 – 63, doi 10.1109/MCOM.2009.5273809.

[31] B. Cummings, T. Zimmerman, B. Robinson, and M. Snyder, Voice over Blue Force Tracking, in Proceedings of *IEEE*

Conference on Military Communications Conference (MILCOM 2006), 23-25 Oct. 2006, Washington, DC, pp. 1 – 5, doi 10.1109/MILCOM.2006.302173.

[32] V.N. Dobrokhodov, I.I. Kaminer, K.D. Jones, and R. Ghabcheloo, Vision-based tracking and Motion Estimation for Moving targets using Small UAVs, *Proceedings of the 2006 American Control Conference Minneapolis*, June 14 - 16, Minnesota, USA, pp. 1428 – 1433, doi 10.1109/ACC.2006.16564418

[33] M. Joze, M. Fras, and Z. Cucej, New approach to the modeling of Command and Control Information Systems, in Proceedings of *IEEE Conference on Military Communications (MILCOM2008)*, 16-19 Nov. 2008, San Diego, CA, pp. 1 – 7, doi 10.1109/MILCOM.2008.4753134.

[34] H. Chang, and J. Zhang, New metrics for clutter affecting human target acquisition, *IEEE Transactions on Aerospace and Electronic Systems, vol 42, issue 1*, Jan. 2006, pp. 361 – 368, doi 10.1109/TAES.2006.1603429.

[35] T. Saarelainen, and J. Jormakka, Interfacing Collaboration and Command Tools for Crises Management Military Command and Control Systems, International Journal of Electronic security and Digital Forensics, vol 3, No. 3, 2010, pp. 249 – 264.

[36] D. Damian, S. Marczak, M. Dascalu, M., Heiss, and A. Liche, Using a Real-Time Conferencing Tool in Distributed Collaboration: An Experience Report from Siemens IT Solutions and Services,IN Proceedings of *Fourth IEEE International Conference on Global Software Engineering (ICGSE 2009)*, 13-16 July 2009, Limerick, pp. 239 – 243, doi 10.1109/ICGSE.2009.31.

[37] A. Martini, S. Mourão, and W. Silva, What Now: A System to Enable Videostream in a Mobile Network, in Proceedings of *IEEE Conference on Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, 2009 (COMPUTATIONWORLD '09)*, Athens, Greece, pp. 557 – 562, doi 10.1109/ComputationWorld.2009.99.

[38] T. Saarelainen, and J. Timonen, Tactical Management in near real-time Systems, in Proceedings of *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (Cogsima 2011)*, CogSIMA2011, Miami Beach, Fl, U.S.A., pp. 240 - 247, doi 10.1109/COGSIMA.2011.5753452.

[39] J.M. Comstock, C.D. Freeborg, B.J. Martinez, C.P. Turner, D.L. Wilson, and T. Rippert, Future Force Warrior, in Proceedings of *IEEE Conference on Systems and Information Engineering Design Symposium*, pp. 234 – 238, doi 10.1109/SIEDS.2006.278682.

[40] C.K.S. Leung, and Q.L. Khan, Distributed Mining of Constrained Patterns from Wireless Sensor Data, in Proceedings of *Web Intelligence and Intelligent Agent Technology Workshops (WI-IATW2006), 2006*, pp. 248 – 251, doi 10.1109/WI-IATW.2006.59.

[41] A.M. Mahdy, and D.S. Deogun, Optimizing Free Space Optics for City-Wide Wireless Networks, in Proceedings of *the Sixth International Conference on Networking (ICN '07)*, Martinique, pp. 66, doi10.1109.2007.66.

[42] R.S. Raghavan, A. Kam, and R.Y. Mannepalli, Modeling & simulation to study the performance of hybrid free space optical/rf military communication networks, in Proceedings of *IEEE Conference on Military Communications Conference (MILCOM 2008)*, 16-19 Nov. 2008, San Diego, CA, pp.1 – 7, doi 10.1109/MILCOM.2008.4753149.

Exploiting Model Variability in *ABS* to Verify Distributed Algorithms

Wolfgang Leister
Norsk Regnesentral
Oslo, Norway
wolfgang.leister@nr.no

Joakim Bjørk and Rudolf Schlatte and Einar Broch Johnsen
Institute of Informatics, University of Oslo
Oslo, Norway
{joakimbj,rschlatte,einarj}@ifi.uio.no

Andreas Griesmayer
Department of Computing,
Imperial College London, UK
a.griesmayer@imperial.ac.uk

***Abstract*—We show a way to evaluate functional properties of distributed algorithms by the example of the AODV algorithm in sensor networks, *Creol* and *ABS* models, and component testing. We present a new method to structure the evaluation work into the categories of techniques, perspectives, arrangements, and properties using executable models. We demonstrate how to use this structure for network simulations and component testing using *Creol* models and demonstrate how the delta modelling technique of the *ABS* language can be used to facilitate the approach.**

***Keywords*—formal analysis, modelling, model checking, testing, routing algorithms.**

I. Introduction

With increasing miniaturisation of hardware on one hand, and reduced production cost and power consumption on the other, computational devices are becoming virtually omnipresent and pose new challenges in software development. A novel systematic methodology for verification of such distributed system was presented previously [1] on the example of wireless sensor networks (WSN) [2] modelled in the executable modelling language *Creol*. In this paper, we extend this work by giving more details on the verification process and reporting on advancements in modelling by using the Abstract Behavioural Specification (*ABS*), a recently developed successor of the *Creol* modelling language [3].

The sensor network of our case study consists of spatially distributed autonomous sensor nodes that communicate using radio connections. Each node can sense, process, send, and receive data. We concentrate on the verification of a *distributed algorithm* for ad-hoc networks between the sensor nodes to route data packets of the participating nodes. There are many functional and non-functional requirements for WSN: routing must fulfil properties for quality of service (QoS), timing, delay, and network throughput; furthermore, we are interested in properties like mobility and resource consumption. When evaluating WSN, autonomous behaviour of the nodes leads to state space explosion during model checking, making evaluation a complex task that requires a combination of techniques from different verification approaches.

The presented structured methodology to verify distributed algorithms introduces the categories of *techniques*, *perspectives*, *arrangements*, and *properties*. This structure is combined with techniques from simulation, testing, and model checking to create a new, unified method for verification of distributed systems. We demonstrate the approach by evaluating a large set of properties on a network using the *Ad hoc On Demand Distance Vector* (AODV) routing algorithm [4].

We detail the modelling process using the new language *ABS*, a successor to *Creol* with a variety of improved characteristics and features. While the basic structure of the models remains the same, *ABS* allows to employ one single executable model that is suitable for simulation, testing, and model checking without the need to develop separate models for each task. In contrast to the models developed in our previous work [1, 5], the new models employ techniques from *software product line* modelling [6] to structure the executable model with a resulting reduction in model size of more than fifty percent.

The remainder of this paper is organised as follows: After introducing the concept of model variability in *ABS* and presenting the used languages and related work (Section II), we discuss the AODV model developed previously and contrast it with the newly-developed model (Section III). Next we present our categories for the validation process (Section IV), present results from network simulation and component testing (Section V), and conclude in Section VI.

II. Abstract Behavioural Modelling with Variability

Diversity poses a central challenge in modern software development. Typically, engineers create different system variants to address a number of concerns ranging from different application contexts to customer requirements [7]. Model-centric approaches to system development that provide an abstract representation of system structure and behaviour are rapidly gaining popularity. There is a lot of research involving feature description languages [8], architectural languages for components [9], the Unified Modeling Language (UML) [10], and state machine-based notations [11–13]. Development processes such as software product line engineering [6] distinguish between generic artifacts that are common to different system variants and product-level system development; these processes are specifically designed to use (and reuse) high-level artifacts. For this reason, software product line engineering is a promising approach to model system diversity. However, a prerequisite for ensuring the consistency

of different views during software product line engineering is a uniform semantic foundation [7].

In our model of routing and forwarding algorithms in wireless sensor networks, we encountered model variability for a variety of reasons; these include different features of the model (timed vs. untimed, adjustable message loss, different routing protocols and sensor layouts) and code adjustments for testing purposes. In the *Creol* version of the model, a preprocessor-based solution for feature modelling was employed [14] to address model variability; in the *ABS* model, we used a more principled approach based on feature modelling and deltas.

The rest of the section presents in more detail the modelling approach of *Creol* and *ABS* and about product line modelling in general.

A. The Creol language

The previous paper [1] used *Creol* to model the AODV algorithm. *Creol* [15, 16] is an *object-oriented* modelling language that provides an abstract, executable model of the implementation of components. The *Creol* tools are part of the *Credo* tool suite [17] that unifies several simulation and model checking tools. The *Credo* tools support integrated modelling of different aspects of highly re-configurable distributed systems both structural changes of a network and changes in the components and offer formalisms, languages, and tools to describe properties of the model in different levels of detail. These formalisms include various types of automata, procedural, and object-oriented approaches.

To model components, *Creol* provides behavioural interfaces to specify inter-component communication. We use intra-component interfaces together with the behavioural interfaces to derive test specifications to check for conformance between the behavioural model and the *Creol* implementation. Types are separated from classes, and (behavioural) interfaces are used to type objects.

Creol objects can have active behaviour. They are concurrent, so that, conceptually, each object encapsulates its own processor. Each method call in *Creol* results in the creation of a new process. This means that the calling process continues to run and receives the result of the method call later using a *Future variable* [18]. Since all object fields are private, processes running on different objects can run in parallel safely, without the need for locking of data structures.

Scheduling *within an object* is based on explicit processor release points, i.e., processes within an object cooperate on scheduling. This cooperative scheduling makes it not only possible to reason about and prove the correctness of parallel programs (since scheduling points are apparent in the program text), but, in our experience, it makes code also easy to read and understand with confidence.

During object creation, a designated run method is automatically invoked if present in the class definition; this method provides active object behaviour.

Creol includes a compiler, a type-checker, and a simulation platform based on Maude [19], which allow simulation, guided simulation, model testing, and model checking.

TABLE I
ABS LANGUAGE LAYERS AND THEIR ROLE IN MODELLING SYSTEM DIVERSITY.

Language layer	**Modelling role**
Functional layer	Specifies internal computations in behavioural modules
Concurrent object layer	Specifies communication and synchronisation of behavioural modules
Delta layer	Modifications to core behavioural modules
Product line configuration layer	Links features to sets of delta modules
Product selection layer	Selects features and initialises a product

Creol only provides limited features for expressing diversity through the use of a pre-processor in the extension *CreolE* [14]. This approach has the disadvantage that the model can become hard to read since all possible aspects are present side-to-side, obfuscating the flow of control. Other features not available in *Creol* include user-defined data types and user-defined functions. Hence, a successor to *Creol* was developed to address these deficiencies.

B. Abstract Behavioural Specification

The *Abstract Behavioural Specification* (*ABS*) [3, 20] modelling language and its accompanying tool framework proposes an approach to the engineering of system diversity that provides a uniform semantic foundation. It supports the precise modelling of behaviour for highly configurable, distributed systems in an end-to-end manner. This means that not only the (concurrent) implementation of features is captured, but also the feature space and the dependencies among them. The *ABS* modelling language aims to fill the gap between structural high-level modelling languages, such as UML, and implementation-close formalisms, including programming languages [20]. Furthermore, *ABS* supports the explicit modelling of time-dependent behaviour for object-based systems by means of its real-time extension [21].

ABS has at its core a state-of-the-art, strongly typed, abstract, concurrent, object-based modelling language [3] that is fully executable. *ABS* offers a number of layers to the system engineer. These layers provide a separation of concerns between different aspects of a system model. Table I shows the different language layers provided in *ABS* and their role in the system modelling. We now explain the purpose and particular features of each layer.

The functional layer of ABS is used to provide a model of internal computation that abstracts from low-level implementation details such as the imperative representation of data structures. This layer consists of user-defined parametric algebraic data types and parametric functions over the terms of these types, including pattern matching. The integration of the functional layer into the object-oriented models results in a very succinct representation of internal computation in the objects that allows the engineer to focus on the communication

and diversity aspects in the upper layers of the model without abstracting from data flow in the model. Thus, *ABS* models are abstract, yet faithful to the data and control flow of the target systems. The *ABS* functional layer has no counterpart in the earlier *Creol* language, which relied on a fixed set of data types and had no user-defined functions.

The concurrent object layer of ABS is used to define the modelling artifacts that represent the system entities in terms of concurrent object groups. Since representation objects can be replaced by terms from the algebraic data types of the functional layer, the engineer may abstract from most representation objects in the model. As a consequence, objects in *ABS* are fairly high-level entities, which are more similar to actors [22, 23] than to Java objects. The concurrent object layer is based on asynchronous method calls between concurrent objects, decoupling communication and synchronisation in the models.Shared memory and synchronous method calls are only permitted among closely collaborating synchronous groups of objects in *ABS*. Otherwise, objects communicate asynchronously and use message passing to update the state. Asynchronous method calls do not transfer control between the caller and the callee. Instead, the reply to a method call may be retrieved by synchronising on a *future* [18].Inside the concurrent object groups, *ABS* uses collaborative scheduling to provide reasoning control in the interleaving of active and reactive behaviour: a method activation can only be suspended by explicitly yielding control. If no method is active, any enabled method activation may proceed. The concurrent object layer is a straightforward extension of the *Creol* object model, keeping most of the semantics and introducing concurrent object groups.

The purpose of the remaining layers in *ABS* is to engineer system diversity. None of these layers are present in *Creol*. While *ABS* is an object-based language and, hence, compatible with the UML world, code reuse by inheritance, which tends to be brittle, is excluded. Instead, system diversity in *ABS* is captured by *delta modelling* (e.g., [24]), which represents a set of systems by a designated core system and a set of system deltas specifying modifications to the core system. Delta modelling is an incremental composition technique for structured diversity that is highly compatible with feature-oriented software development [25] and also a good match for agile and evolutionary development approaches [26].

The delta layer of ABS is used to specify structured changes to the set of classes by adding or removing variables or methods from a class or by redefining methods in the class. A delta consists of a set of such changes, and may additionally add or remove classes. The integration of deltas into a (core) model happens in a given order at compile time, transforming the model. *The product line configuration layer of ABS* is used to define features in the software product line as sets of deltas. Thus, selecting a feature consists of applying the deltas in the corresponding set in a given order. *The product selection layer of ABS* provides the means to configure a system in order to obtain a given variant of the product line. This is done by selecting the features that should be provided in a specific product of the product line.

C. Timed Modelling

To simulate and verify functional correctness of models as well as timing- and performance-related criteria, a notion of time needs to be included in the model. A timed extension for *ABS* has been developed in Bjørk et al. [27]. In its present form, *timed ABS* employs discrete time and run-to-completion semantics, i.e., the maximum amount of computation is performed in each time interval before the (simulated) clock is advanced. Timing behaviour of models is explicitly encoded, thus being visible to the modeller.

The new language elements are:

- A statement `duration(`*`best`*`, `*`worst`*`)` that blocks the current object between *best* and *worst* time units where no other process can execute on that object. This can be used to model CPU-intensive tasks.
- A new guard condition: the statement `await duration(`*`best`*`, `*`worst`*`)` causes the current process to be suspended between *best* and *worst* time units, letting other processes of the current object execute in the meantime. This can be used to abstractly model the timing behaviour of interactions with external systems, such as interactions with an external database that is not explicitly modelled.
- A function, `now()`, that returns the current model time as a monotonically increasing integer value. In practice, this function can be used for recording completion times and bookkeeping.

Using `duration` and `await duration`, the modeller can express, e.g., message transfer delays in the network and processing delays in the sensor nodes for the AODV model.

D. Related Work

Showing functional correctness and non-functional properties for algorithms employed for WSN helps the developers in their technical choices. Developers use a variety of tools, including measurements on real implementations, simulation, and model-checking. When developing algorithms for packet forwarding in a WSN, simulation results must be compared with the behaviour of known algorithms to get a result approved [28]. Approaches using simulation, testing, and model checking during the development process use one or more of the following: modelling, traces, runtime monitoring by integrating checking software into the code (instrumentation) [29], or generating software from models automatically [30].

Simulation systems are used to analyse performance parameters of communication networks, such as latency, packet loss rate, network throughput, and other metrics. Most of these systems use discrete event simulation. Examples for such simulation systems [31] include OPNET [32], OMNeT++ [33], the network simulator ns-2 [34], its successor ns-3 [35], or mathematical frameworks like MathWorks [36]. Many of these tools have specialised libraries for certain properties, hardware, and network types. While these network simulators are often used to evaluate the performance of algorithms, they are

primarily not designed for model checking tasks. This implies that code for model checking needs to be implemented in these simulators rather than using these features as integrated parts.

The CMC model checker [29] has been applied on existing implementations of AODV by checking an invariant expressing the loop-freeness property. In that work, both specification and implementation errors were found and later corrected in more recent versions of both specification and implementations. CMC interfaces C-programs directly by replacing procedure calls with model-checker code, thus avoiding the need to model AODV. Wibling et al. use the model checking tools SPIN and *UPPAAL* to verify properties for the correct operation of ad hoc routing protocols [37], such as the LUNAR and DSR algorithms [38]. They use *Propagation Localised Broadcasting with Dampening* (PLBD) as a basic operation, and perform model checking on the LUNAR and DSR algorithms. Both LUNAR and DSR are related to AODV, but use different mechanisms. Chiyangwa and Kwiatkowska [39] uncovered in a timing analysis in *UPPAAL* that many AODV connections unnecessarily timed out before a route could be established in large networks. To avoid this, they proposed to set the timeout value dependent of the network diameter.

Timed automata implemented in *UPPAAL* have been used for validating and tuning of temporal configuration parameters and QoS requirements in network models that allow dynamic re-configurations of the network topology by Tschirner et al. [40]. The strength of *UPPAAL* is that both average-case and worst-case behaviours can be analysed. Tschirner et al. compare their results with a similar implementation in OMNeT++. They found that the results from both simulations coincide closely. While the *UPPAAL* implementation is more high-level, the task of implementing the C++ code for OMNeT++ is rather time-consuming.

The model checker *Vereofy* [41, 42], part of the *Credo* tools, was used to analyse aspects of sensor networks and AODV, as presented by Baier et al. [43]. For *Vereofy* formal semantics relies on constraint automata. Thus, a model of a WSN describes the behaviour of the sensor nodes and the network at the interface level. The specification of the interface behaviour of a sensor node is given in terms of *CARML* (Constraint Automata Reactive Module Language) sub-modules for sensing, receiving and sending. For unicast and broadcast the communication media have been modelled as dynamic component connector networks composed with the help of *RSL* (Reo Scripting Language).

While *Vereofy* uses an automaton approach and process algebra with exogenous coordination, *Creol* and *ABS* are based on executable object-oriented models. Note that properties that can be checked by *Creol* or *ABS* are not necessarily suitable to be checked by *Vereofy*, and vice versa. We also used *Vereofy* as a reference for evaluating the properties and as source for the traces employed for the component testing.

Real-Time Maude [44] is a language and tool supporting the formal specification and analysis of real-time and hybrid systems, based on rewriting logic. It is particularly suitable to specify object-oriented real-time systems. Real-Time Maude can be seen as complementing on the one hand timed/hybrid automaton-based tools such as *UPPAAL*, HyTech, and Kronos, as well as, e.g., timed Petri nets. The OGDC-algorithm used in certain sensor networks has been simulated and model-checked in Real-Time Maude [45]. The comparison of these simulation results in Real-Time Maude against simulation results in ns-2 have uncovered weaknesses in a concrete ns-2 simulation.

III. Modelling the Components and the Routing Algorithm

Distributed applications can be described in terms of components interacting in an open environment based on the mechanisms of *Creol* [46]. This framework models components and the communication between these components, and executes the models in rewriting logic. Different communication patterns, communication properties, and a notion of time are supported. The lower communication layers use tight, loose, and wireless links.

Based on this work, we defined a model of AODV in a WSN using *Creol* [47] that expresses each node and the network as objects with an inner behaviour. The interfaces of the objects describe the communication between the nodes and the network object. In Figure 1, we show the object structure of the model, including the most important interfaces of a node. Note that the object structure shown in Figure 1 is rather generic and can be employed for modelling AODV in other simulators. This interface model is in contrast to the interface model used by the test harness shown later in Figure 4.

Inside a node, its behaviour was implemented in *Creol* as routines that are not unlike real-world implementations. The model contained different aspects (message loss, timed simulation, different routing algorithms) that were enabled or disabled via a pre-processor.

In the remainder of this section we discuss an *ABS* model that has been implemented later with the same interfaces using the experiences from the previous *Creol* model and using the advantages of *ABS*.

The transition from *Creol* to *ABS* reduced the size of the model by about 50%, as measured by line count. Much of the reduction comes from the new language features of *ABS*, predominantly the added functional layer. In addition to being shorter, the new code is also more readable. The model is structured into a basic layer with a number of deltas adding additional features. The basic layer specifies the interfaces and object structures. It implements a simple flooding protocol without routing or retransmission. Based on that simple but working model, a delta replaces the routing functionality in the sensor class with new code that implements the AODV protocol. This layer also adds attributes for storing the local routing tables in each sensor object. Additional deltas implement message timeout and message loss simulation facilities. In the *Creol* model, all these functionalities were implemented in one place and switched on and off by preprocessor directives; the new structure allows to implement, read, and understand the different aspects in isolation.

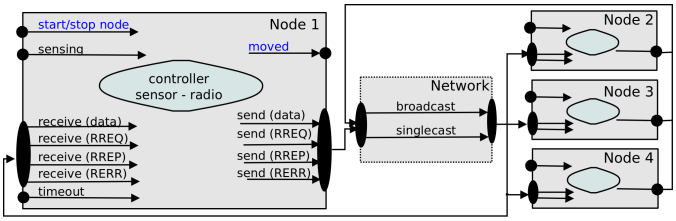


Figure 1. Objects of a WSN model and their communication interfaces.

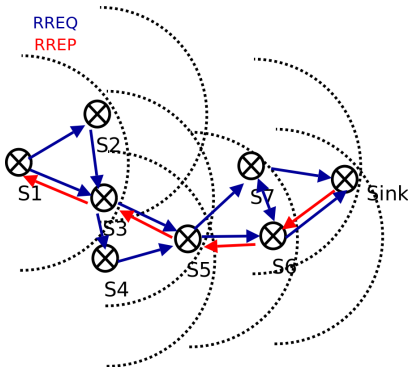


Figure 2. Example of AODV message propagation in a WSN with eight nodes.

A. The AODV Routing Algorithm

The purpose of a routing algorithm is to establish a path between a source node and a sink node, so that data can flow from the source node to the sink node via forwarding nodes in-between. AODV is a reactive routing protocol that builds up the entries in the dynamic routing tables of nodes only if needed. AODV can handle network dynamics, e.g., varying wireless link qualities, packet losses, and changing network topologies.

When a node wants to send a message to a sink node and the next hop cannot be retrieved from the local routing table, it initiates a route discovery procedure by broadcasting RREQ (route request) messages. Nodes that receive a RREQ message will either send a RREP (route reply) message to the node that originated the RREQ message if the route is known; otherwise the node will re-broadcast the RREQ message. This procedure continues until the RREQ message reaches a node that has a valid route to the destination node. The RREP message is unicast to the source node through multi-hop communications; as the RREP message propagates, all the intermediate nodes also establish routes to the destination. After the source node has received the RREP message, a route to the destination has been established, and data packages can be sent along this route. As an illustration, Figure 2 shows an example of a small WSN with eight nodes, where the potential RREQ messages are shown in blue, while the RREP messages are shown in red. Note that other paths for the RREP messages are possible in this example.

The essential entries of the routing table in each node include the next hop, a sequence number, and the hop count to the sink node. The hop count is the most common metric for routing to choose between routes when multiple routes exist. The sequence number is a measure of the freshness of a route.

When communication failures imply a broken route, the node that is unable to forward a message will inform other nodes so that the routing tables can be updated. To do this, it sends a RERR (route error) message along the reverse route that is also stored in the nodes. Thus the source node will become aware of the broken route, and initiate a new route discovery procedure.

B. Modelling Message Types

For modelling AODV messages and their operations, we use the functional layer of *ABS*. For simplicity, object references to the sending, origin, and target nodes serve as tokens inside messages. The message payload is represented by a single integer since we do not model higher-level protocols that process message data.

```
data MsgType = RREQ | RREP | RERR | PAYLOAD;
data Message = Msg(MsgType msgType, Node sender,
   Node origin, Node target, Int originSeqNo,
   Int ttl, Int hops, Int content);
```

These data type definitions also generate functions (`msgType()`, `sender()`, etc.) to access the components of a message.

C. Modelling Active Objects

The `Node` type used above is an *interface type*. In *ABS*, classes are not types, so each class must implement at least one interface if it wants to be assignable to a variable. Note that classes without interfaces can be meaningful if they have active behaviour – objects of that class can interact with the rest of the model via references passed to their constructor.

```
interface Node{
   Unit receiveMsg(Message msg);
   Unit timeout(Message msg);
}
interface Sink extends Node {}
interface Sensor extends Node {
   Unit start();
}
```

The methods `receive`, `timeout`, and `start` implement the component behaviour shown in Figure 1. Sending the messages and the network behaviour are implemented by the following interface:

```
interface Network{
   Unit createLink(Node node1, Node node2);
   Unit deleteLink(Node node1, Node node2);
   Unit send(Message msg);
}
```

Besides the method `send()`, the `Network` interface contains the methods `createLink` and `deleteLink` that are used for configuring the network topology and expressing which nodes are neighbours. Neighbours can directly receive each other's messages.

D. Implementing and Augmenting Classes

The class that describes a sensor node is defined as follows:

```
class SensorNode(Network n, Sink s) implements Sensor
{
  ...
  Unit receiveMsg(Message msg) {
    msg = incHopCount(msg);
    this.recordMsg(msg);
    this.forwardMsg(msg);
  }
  ...
}
```

Most of the implementation is elided for brevity, but we show an implementation of `receiveMsg`, which is called by the network when a neighbouring node sends a message. To explain the model variability principles of *ABS*, we first present code that implements a *flooding* protocol where routing is not involved. In this code, the method `forwardMsg` checks whether the incoming message already has been seen; otherwise, it calls *n.send()* for retransmission.

To add the AODV routing protocol to the model, we use a *delta*. This allows us to selectively modify classes, adding and replacing methods and member variables. To add AODV routing, the delta replaces the method `receiveMsg` with a message that also updates the routing table before invoking the original method via the `original()` call.

```
delta AODV {
  modifies class SensorNode{
    modifies Unit receiveMsg(Message msg){
      this.updRoutingTable(msg);
      original();
  }
    ...
  }
}
```

The delta also adds member variables for the routing table, and modifies `forwardMsg` with functionality to handle the different kinds of messages and only re-send messages if the node is on the path to their destination. Again, these parts are elided for brevity.

In the end, when running simulations, we can choose which model to run by defining a product containing the right combination of features, and giving its name as parameter to the *ABS* tool chain.

```
product FloodingBSN(Flooding);
product AODVBSN(AODV);
```

We refer to the work by Schaefer et al. [24] for details about delta modelling.

E. Comparison to the Previous Creol Model

The *ABS* language has been designed from the experiences of *Creol*. We outline some improvements of *ABS*. In *Creol*, only basic data types and classes are available for modelling purposes.

In a first attempt to model AODV in *Creol*, we implemented messages as objects. However, this soon caused an overload of the interpreter in the sense of high execution time and large space requirements, making this model rather impractical. Moreover, since messages do not have an independent behaviour, the use of objects for messages is not required.

Modelling messages as integer numbers would theoretically be possible, but does not allow to add annotations in the form of log messages to the object. Since the underlying Maude interpreter does not allow for writing log files, we need another mechanism for creating traces of messages. From a practical perspective, it is valuable for evaluations to follow the path of messages after having performed a simulation. The final implementation of the *Creol* model, therefore, used maps of strings, where annotations could be added to the message. This allows the extraction of the wanted information from the Maude state file. Modelling messages in this way proved feasible for simulation purposes, but manipulating messages was unwieldy and contributed to the comparatively larger code size of the *Creol* model.

The *Creol* nondeterministic choice operator (`[]`) was used to model message loss. While this was a succinct formulation of the semantics of message loss, it was not really suitable for simulation purposes because we could not adjust the likelihood of message loss. Consequently, *ABS* introduced a proper random function that was used for expressing a parameterisable likelihood of message loss, allowing for Monte Carlo simulation. Despite the use of (pseudo-)random behaviour, simulation results are still reproducible because the seed value for the random number generator in Maude can be supplied as a parameter.

As already mentioned, the *Creol* model mixed different model parameters in-line in the code, relying on a preprocessor to generate the desired code. While this approach works, the resulting model is hard to follow for the reader. The product line- and delta-modelling capabilities of *ABS* proved to be a good approach to extract different behavioural aspects of the model into their own semantic units and generate models with the desired aspects on demand.

IV. Methodology for Simulation, Component Testing, and Model Checking

In this section, we show how to evaluate and validate the functional behaviour of the AODV model in the *Credo* framework [17]. While the *Credo* tool set is based on *Creol*, it is easy to see from the previous discussions that the presented concepts are applicable to the *ABS* model presented in the previous sections. We present the *techniques*, *perspectives*, *arrangements*, and *properties* necessary for the validation and show how to evaluate selected non-functional properties.

A. Techniques for Simulation, Testing, and Model Checking

In order to evaluate the properties of a model, several *techniques* are used to provide the necessary technical measures and procedures to make a model amenable to verification. In general, the following modifications can be applied to the model in preparation for simulation, testing, and model-checking:

Auxiliary variables are added to the model to improve the visibility of a model's behaviour. They must not alter the behaviour and are updated when certain relevant events happen, e.g., a counter is incremented when a new instance is created. When running a simulation, these values can be extracted from the state information and visualised in a step-by-step execution or after the execution of the model execution terminates.

Assertions might be necessary depending on the functional requirements to check. While a number of properties can be checked at the final state using auxiliary variables, properties on the transient behaviour of the model require a check during runtime. For such cases, *Creol* provides *assertions* that stop the execution of a model when the condition is violated. The state that caused the violation of the property is then shown for further analysis.

Monitors are pieces of software that run in parallel to the actual model and are used for properties that go beyond simple assertions. A monitor constitutes an automaton that follows the behaviour of the model to decide the validity of a path.

Guarded execution replaces nondeterministic decisions by calls to a guarding object, here denoted as the *DeuxExMachina* module. This allows to check the behaviour of the model under different conditions, while still maintaining reproducibility of the runs. This technique also specifies certain parameters of the environment, like failure rates of the network.

Fault injection adds a misbehaving node (possibly after a certain time) to check error recovery properties. For instance, misbehaviour in a node may be triggered when energy is used up. Such behaviour can be implemented by sub-classing nodes and implementing certain misbehaving routines in the subclass.

Property search employs model checking techniques to check whether certain conditions hold for all or a given subset of states. Such a search can be directly performed by *Maude*, the execution engine for our interpreter, without the need of implementing the search code in the model.

B. Perspectives

A *perspective* describes the scope of an evaluation. For the AODV model, we developed two perspectives: (*a*) observing the behaviour of the entire network configuration including all nodes and the network and (*b*) observing the behaviour of one node. Testing, simulation, and model checking can be performed from different perspectives and levels of detail for a given model. For AODV, a holistic perspective focuses on the networking aspect of the nodes implementing all the involved nodes and the network in one model. However, for model checking such a model leads to a high number of states and long execution time. Therefore, for realistic models the networking perspective is not feasible.

For the perspective of testing a single node, we use the same model code for the nodes in the holistic perspective, but instantiate only one node explicitly. The network is replaced by a *test harness* that impersonates the network and the remaining nodes. The behaviour and responses of the test harness are determined by a rule set that is derived from traced messages between the nodes, as outlined in Section V-B.

C. Arrangements

An *arrangement* denotes a set of configuration settings that influences how the model operates. Examples are the use of untimed or timed models, changes to the node topology, perfect or unreliable communication, communication failures, timeouts, and energy consumption. *ABS* supports different arrangements natively using its delta layer, while for *Creol* a preprocessor can be used. Examples for arrangement entities that can be selected in the models, together with implementation details for the AODV model, are given in the following:

The *communication behaviour* in our model can be set to be either reliable, non-deterministic, or one of several packet loss patterns including random packet loss. (Note that pure non-deterministic behaviour in a simulation currently is not useful due to restrictions in the implementation of the underlying runtime system, and in general because of non-reproducibility of simulation results.) Using the differences in communication behaviour we can study how the algorithm behaves when communication packets can get lost.

Topology changes are used to check the robustness of the protocol. They can be triggered by certain events, e.g., after a certain number of messages or after a certain amount of time for timed models. A topology change affects the connection matrix in the network and triggers the AODV algorithm to find new routes in the model.

The *timed model* is realized using discrete time steps and introducing a global clock in the network object and internal clocks in the nodes that are synchronised when a task is performed in one or more nodes. This allows, e.g., to reason about messages being sent simultaneously, which eventually will lead to packet loss. Also, the effect of collisions can be shown without using non-deterministic packet loss. The use of a timed model is most viable together with topology changes since the topology needs to be re-installed for a state when another branch is searched in model checking.

Energy consumption is modelled using an auxiliary variable in each sensor node with an initial amount of energy. For each operation, a certain amount of energy is subtracted until the capacity is too low to perform operations on the radio. This indicates a malfunction of the sensor node. Such a node does not perform any actions and represents a topology change of the network, since given path are no longer valid. This allows us to identify in which cases an energy-restricted network can perform communication and whether AODV can find routes around an energy-empty node.

Note that arrangements for memory and buffer sizes can be implemented similarly. When maximum memory size is reached, a node will alter its behaviour and stop performing certain actions.

Timeouts are modelled nondeterministically by the use of a global guarding object and can occur between a message is sent and the corresponding answer is received. AODV employs

timeouts in order to work in environments where communication errors can occur and sends messages repeatedly in case an expected reply has not been received from the network.

D. Properties

A *functional property* is a concrete condition that can be checked for given arrangements, while non-functional properties are values given by metrics. For AODV, we chose the following functional properties: *a*) correct-operation, *b*) loop-freeness, *c*) single-sensor challenge-response properties, *d*) shortest-path, *e*) deadlock-freeness (both for node and for protocol), *f*) miscellaneous composed system properties, and additionally some non-functional properties.

Correct-operation: For a routing algorithm to be correct, it must find a path if a path exists, i.e., it is *valid* for some duration longer than what is required to set up a route from sender to receiver [37, 38]. Checking this property requires the algorithm-independent predicate whether a route exists. In the absence of topology changes, this predicate can be calculated beforehand. When topology changes are possible, however, we need to check the existence of a path between sender and receiver at any step in the algorithm. Since checking this property in *Creol* involves explicitly visiting all nodes, this increases the reachable state-space of the model. To evaluate this predicate effectively, a suitable implementation would be to interface a *Maude* function, which is possible in *ABS*. Meseguer and Rosu [48] give an overview of existing work on model-checking language semantics in Maude, which can be used for *Creol* and *ABS*.

A related property to evaluate is whether a route is re-established after a transmission error given a path still exists. We also evaluate how long the path is interrupted after a transmission error occurs.

Loop-freeness: A routing loop is a situation where the entries in the routing tables form a circular path, thus preventing packets from reaching the destination. The invariant for loop-freeness [29] of AODV must be valid for all nodes. It uses *sequence numbers* of adjacent nodes, and the number of hops in the routing tables as input. The loop-freeness property is checked every time a message is transmitted between nodes. To do this the network-object calls a routine that checks the loop-freeness invariant in an assertion. Since this assertion is complex and contains nested loops, it again should be implemented as a call to a *Maude* function instead of *Creol* or *ABS* code.

Single-sensor challenge-response: The reaction of one node under test is checked using component testing (Section V-B). Messages are sent to the node under test, and the responses from this node are matched against all correct responses. The correct responses are extracted from specifications or from running simulations using different implementations. The single-sensor properties that can be checked express a certain behaviour or the absence of a certain behaviour after a challenge, e.g., whether an incoming message leads to a specified state change in the node or whether the node sends an expected response messages.

Shortest-path: Here, we investigate whether the AODV algorithm finds the shortest path for the paths between the source and sink node; also other metrics for paths could be checked. While AODV finds the shortest path in the case of no packet loss, it does not always fulfil the shortest-path property in the case of packet loss. To check this property we count the number of hops that each payload-message takes from the source to the sink and compare it with the shortest existing path between the source and sink.

Deadlock freedom: Deadlocks in a node, in the protocol or in the model are a threat to robustness, and can reveal errors in the specification, implementation, or model. Global deadlocks will automatically be detected by the underlying Maude implementation and result in an inspectable error state of the model.

Miscellaneous composed-system properties: Examples are properties that state whether valid routes stay valid, avoidance of useless RREQ messages, number of messages received, timing properties, and network connectivity. Most of these are implemented by adding counter variables and predicates.

Non-Functional Properties: Non-functional properties from the application domain such as timing, throughput, delivery ratio, network connectivity, energy consumption, memory and buffer sizes, properties of the wireless channel, interferences, mobility, or other QoS properties can be evaluated by using counter variables and additional code for the model. Note that for most non-functional properties the use of *Creol* might not be the optimal choice, since it is best suited for the evaluation of conformance or violation of non-functional properties. *ABS* has some support for modelling deployment scenarios and resource consumption [49, 50], which could be extended to cater to our modelling requirements.

V. Holistic and Component Testing

Instrumented *Creol* and *ABS* models can be used for different verification and testing techniques: symbolic simulation, guarded test case execution, and model checking. Auxiliary code for assertions and monitor state is added and executed together with the model code. This increases the size of the states and therefore poses a handicap for model checking. A more light weight approach would implement the monitors directly within the checking tools. This is, however, not yet available in the analysis tools. All experiments in this section were first performed on the *Creol* models. TABLE II gives a detailed list of properties that were identified as being of interest for the network example with the symbol • marking properties that were evaluated, ⊙ partially evaluated and ○ not evaluated yet for simulation and testing. Model checking was used for generating execution traces for testing as detailed in Section V-B. The *Creol* models have been re-implemented in part to *ABS*. Unsurprisingly, these show the same behaviour while the model size, measured in lines of code, is reduced.

A. Holistic Testing

For our evaluation of the network properties we used simulation using techniques such as auxiliary variables and

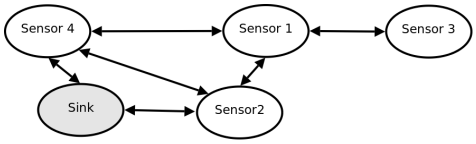


Figure 3. The network used as example in our simulations.

assertions. Most of our experiments used a network with symmetrical communication via four sensor nodes and one sink node, as shown in Figure 3. We also experimented with models of 6, 15, and 30 nodes for selected arrangements. The evaluated model consists of code for the network nodes as well as an explicitly modelled network that transmits the messages between the nodes such that, e.g., a broadcast message of *Sensor 2* reaches the nodes *Sink*, *Sensor 1*, and *Sensor 4*. This gives the flexibility to simulate the AODV model using various arrangements including reliable networks, lossy networks, timeouts, energy consumption, and timed modelling.

Symmetrical communication means that whenever a node *A* can transmit a message to node *B*, then communication in the reverse direction can also take place. By changes to the network structure, we also could show that this is a property the AODV routing algorithm in fact relies on. This is because if node *B* receives a broadcast message from *A*, it updates is routing entry for messages destined for *A*. We also checked selected properties from the classes (a), (b), (d), (e), and (f) presented in Section IV-D for the composed network.

Reliable communication: As long as the network is connected, the evaluations showed that the modelled AODV algorithm fulfils the properties (a), (b), (d), (e), and (f) of Section IV-D. We emphasised on the evaluation of packet loss, and loop-freeness assertion. Other predicates for loop-freeness were also used (which failed as expected), and small, faulty changes in the model were introduced (which led to expected failures of the loop-freeness property). The shortest path property was fulfilled in all simulated occasions.

Lossy communication: When simulating lossy communication both for singlecast and for broadcast messages, the packet loss rate f.(xxvi) increases as expected. We also observed an increased number of RREQ and RREP messages in the system using auxiliary variables.

Timeouts: The model allows re-sending of lost RREQ messages up to a certain number of times, using a timeout mechanism. We could observe that this mechanism decreased the packet loss rate, f.(xxvi), but at the same time does not prevent all packet loss for payload packets.

Energy consumption: Using the energy consumption arrangement we can force a communication failure of certain nodes after some actions. Using this arrangement we can study the re-routing behaviour in detail, including the packet loss rate f.(xxvi).

Timed model: Using the timed model we can study the number of time steps needed for sending messages, as well as controlling the number of actions being performed simultaneously. We observed that the packet loss rate f.(xxvi) is different to the untimed case, which is expected.

TABLE III
NUMBER OF REWRITES AND RUN-TIME FOR SAMPLE ARRANGEMENTS AND PROPERTIES.

#	t steps	energy	loss	timeout	#rewrites	time
5	500	–	none	never	$9.4 \cdot 10^6$	17.1s
	5000	–	none	never	$62.8 \cdot 10^6$	114.8s
	500	–	10%	never	$10.7 \cdot 10^6$	19.5s
	500	–	10%	1/10	$12.1 \cdot 10^6$	22.3s
	500	50	10%	1/10	$8.3 \cdot 10^6$	15.5s
	untimed	50	10%	1/10	$11.6 \cdot 10^6$	17.9s
	untimed	–	10%	never	$32.5 \cdot 10^6$	14.8s
6	untimed	–	10%	never	$90.5 \cdot 10^6$	40.9s
15	5000	–	none	never	$2.7 \cdot 10^9$	31m
30	5000	–	none	never	$24.8 \cdot 10^9$	8h

Using the timed model, we could observe a model deadlock, e.(xix), which is caused by the way the model is implemented, and certain properties of the current implementation of the *Creol* runtime system. This observation made changes in the model implementation necessary using asynchronous method calls.

The properties f.(xxi), f.(xxiii), and f.(xxiv) could not be evaluated in a satisfactory manner as they require to store all messages during the simulation. Although the properties can be modelled and evaluated in principle, such an arrangement leads to state explosion and exceeds time and memory constraints of our current setting.

The developed *Creol* model was evaluated by using simulation for sample arrangements and properties. The entire model contains about 1600 lines of *Creol* code excluding comments. After compilation, the resulting code size was about 1050 lines of Maude code, depending on the arrangement. We varied the timing behaviour, the energy consumption, the message loss behaviour, and the timeout behaviour of the model as well as the number of nodes. The results for the tested cases considering the number of rewrites, and execution time on an AMD Athlon 64 Dual core processor with 1.8 GHz is shown in TABLE III. The timing behaviour and the number of nodes are the most significant parameters.

While these values may seem high for a simulation system, we emphasise that the purpose of the *Creol* model is to offer one model that is suitable for several perspectives. While the transition from simulation to model checking consists in changing some few Maude statements, the search space during model checking gets combinatorially too high to be viable already for a low number of nodes.

B. Component Testing of One Node

For component testing, we use one node under test with the same code as for holistic testing. However, we replace the network and all the other nodes using a test harness shown in Figure 4. This harness simulates the possible behaviour of the network and the further nodes as visible by the node under test. The output signals of the node under test are connected

TABLE II
PROPERTIES EVALUATED IN *Creol*; NOTE THAT *S* MARKS SIMULATION, WHILE *T* MARKS TESTING AS THE METHOD OF CHOICE.

Property	Description	Evaluation	S	T
a	Correct Operation	yes; for some arrangements.	●	
b	Loop-Freeness	yes	●	
c	Sgl-sensor challenge-resp.	yes		●
c.(i)	always send with own ID	yes, as invariant during other tests		●
c.(ii)	msg leads to valid route	yes (inferred from other tests)		●
c.(iii)	RREQ w/o route⇒RREQ bc.	yes		●
c.(iv)	RREQ for me leads to RREP	yes		●
c.(v)	RREP triggers route to originator	yes		●
c.(vi)	RREP is rebroadcasted	yes		●
c.(vii)	send if route known	yes (no send if route unknown)		●
c.(viii)	routing table integrity	no		○
c.(ix)	all msg for sink	yes		●
c.(x)	processing without receive	yes (during other tests)		●
c.(xi)	increasing sequence number	yes, as invariant during other tests		●
c.(xii)	neighbour update triggers	n/a (not accessible in black-box test)		
c.(xiii)	updates terminate	yes (implicitly during other tests)		●
c.(xiv)	update success	yes (implicitly during other tests)		●
c.(xv)	only one RREQ	n/a (needs timed interpreter)		
c.(xvi)	Rec. in IDLE mode	n/a		
d	Shortest-Path	yes	●	
e	Deadlock-Freeness	partially	⊙	
e.(xvii)	node deadlock	no		
e.(xviii)	protocol deadlock	yes	●	
e.(xix)	model deadlock	yes	●	
f	Misc. Composed-System	yes	●	
f.(xx)	route stays valid	yes	●	
f.(xxi)	only data msg	possible, not done	○	
f.(xxii)	no RERR	yes	●	
f.(xxiii)	no useless RREQ	possible, not done	○	
f.(xxiv)	RREQ triggers RREP	possible, not done	○	
f.(xxv)	# msg.rec.	yes	●	
f.(xxvi)	packet loss	yes	●	
f.(xxvii)	timing properties	partially	⊙	
f.(xxviii)	network connectivity	yes	●	
f.(xxix)	parameter tuning	partially	⊙	

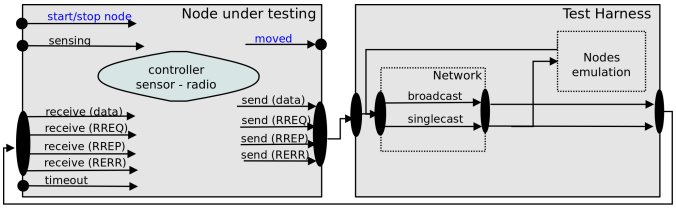


Figure 4. Testing of one node using the network object as a test harness. The test harness replaces the network and all other nodes in the network.

to the test harness, which then in turn generates the inputs that would be received from the real network. The test is then evaluated by studying the output messages of a node when the input messages for the test case are supplied by the harness.

1) Test harness: The task of the test harness is to send messages to the interfaces of the node under test, and to observe its answers. Both input messages and expected answers can be generated from the specification or from traces of real systems or other simulations.

Although incoming broadcast, singlecast, and outgoing packets involve invoking different methods, the *ABS* language, with its object-level parallelism, makes it easy to encode a test case as a single sequential list of statements. Incoming messages are stored in a one-element buffer; the test case simply performs a blocking read on that buffer when waiting for a message from the object under test before sending out the next message to the object. In this way, both creating a test case by hand and generating test cases from recorded traces are feasible.

A test verdict is reached by running the test harness in parallel with the object under test. If the test harness *deadlocks*, it expects a message from the object under test that is not arriving, a test verdict of *Fail* is reached. The other reason for test failure is an incoming message that does not conform to the expectations of the test harness; e.g. by being of the wrong type or having the wrong content.

A test verdict of *Success* is reached if the test harness completes the test case and the object under test conforms to the tester's expectations in all cases.

2) Traces: In addition to domain-specific single-object properties, test cases can be generated from the model implemented with *Vereofy* [43] (see Section II-D). *Vereofy* is a different language and uses independent mechanisms than both *Creol* and *ABS* to create the traces. Therefore, we can compare the two different models against each other, with the *Vereofy* traces as a partial specification for the behaviour of the *Creol* and *ABS* models.

To receive the traces from *Vereofy*, we collect the exchanged data, the content of all variables in the nodes and buffers in the network, and the output of the automaton before and after each step, in addition to the exchanged data. After the state information is removed, we receive a sequence of messages that are exchanged simultaneously.

The messages in the trace are defined using a *Vereofy* struct:

```
TYPE message_t = struct {
  // determine the type of the message
  message_type_t message_type;
  id_t           dest_id;
  // encapsulation of sender and receiver IDs
  address_t      to_ip;
  id_t           from_ip;
  // case 1: sending AODV messages
  hop_counter_t  hop_count;
  seq_no_t       dest_seq_no;
  id_t           orig_id;
  seq_no_t       orig_seq_no;
```

```
  Bool           unknown_seq_no;
  // omit TTL and XFlag for AODV.
  // case 2: for sending data messages
  data_type_t    the_data;
};
```

In the following example, Node 1 sends three RREQ messages to find a route to the Sink (Node 0). The RREP generated by the Sink does not arrive, since it remains in a buffer.

```
{send[1]={RREQ,0,2,1,0,0,1,1,1,data0}}
{receive[0]={RREQ,0,2,1,0,0,1,1,1,data0}}
{send[0]={RREP,1,1,0,0,1,0,0,0,data0},
 send[1]={RREQ,0,2,1,0,0,1,2,1,data0}}
{receive[0]={RREQ,0,2,1,0,0,1,2,1,data0}}
{send[1]={RREQ,0,2,1,0,0,1,3,1,data0}}
```

In the tester for Node 1, written in *ABS*, the run-method waits for the messages denoted as `send[1]`, and sends messages denoted as `receive[1]`. A trace can contain messages that are sent simultaneously, such as the third statement of the above trace. Synchronous communication, the calls of `send[0]={RREP,1,1,...}` and `receicve[1]={RREP,1,1,...}`, take place simultaneously.

Traces received from the node under test are tested against *message patterns*, i.e., we remove details that could lead to spurious test failures not expressing a malfunctioning system. For example, the message sequence number can be chosen by the node, the only requirement is that it be monotonically increasing. This property is checked using an invariant in the tester, but a different concrete message number than that used by the *Vereofy* model will not lead to test failure.

C. Other Mechanisms in the Credo Tools

The *Credo* tools based on *Creol*, *Vereofy*, and *UPPAAL* offer different ways of modelling, supporting different techniques, perspectives, arrangements, and evaluation of properties. The evaluations shown in TABLE II have been performed with a similar model in *Vereofy* for a comparison of results where this was suitable [51]. This shows that *Vereofy* is suitable for Properties c, e, and f.(xx) to f.(xxiv). *Vereofy* offers a model-checking approach based on Reo automata using the exogenous coordination model where the components are represented by their behavioural interfaces. Similar to *Creol* and *ABS*, *Vereofy* supports the verification of components and their communication structure. The concrete case study is described elsewhere [43].

For completeness, we mention that selected properties have been evaluated in *UPPAAL*. However, this model only implemented connectivity between sensors rather than ADOV [40]. Properties in the classes f.(xxv) to f.(xxix) have been verified. We mention, however, that *UPPAAL* is capable for other properties, as the work by Chiyangwa and Kwiatkowska [39] and Wibling et al. [37, 38] shows.

VI. Conclusion

We presented a structured methodology for the evaluation of complex distributed systems by introducing the dimensions of *techniques*, *perspectives*, *arrangements*, and *properties* for this evaluation. We divided the properties used for this evaluation into six property classes, and performed network simulations of the composed system, and component testing of a single node.

Using the network simulation, we evaluated several arrangements. While most of the properties were fulfilled as expected, some properties did not validate in the simulation; this either due to bugs in the model, artificially introduced misbehaviour in the model, properties of the modelled AODV algorithm, or property variants that are not supposed to validate successfully. In one occasion, we could detect deadlocks in the model in a timed-model arrangement, which could be recognised and fixed afterwards. Evaluating other protocols in sensor networks, e.g., proactive dynamic routing protocols, is possible using the variability features of *ABS*, with test cases and test scenarios adapted to these new protocols.

Using component testing, we validated the correct behaviour of a single node against properties extracted from the specification of the AODV algorithm. No deviations from specified component behaviour were identified in this process, which is unsurprising since components had already been extensively used for simulation and animation during initial model development at that point in time. However, the test suite served as an excellent tool in regression testing during subsequent changes and extensions of the model.

Evaluating the properties of the AODV algorithm, we had previously encountered several challenges, such as modelling suitable abstractions, using language constructs of *Creol*, and observing the properties from a suitable perspective. The major challenge when evaluating the AODV algorithm from a network perspective is to avoid a high number of states in the underlying interpreter. We showed how to overcome modelling difficulties in the existing *Creol* model by refactoring it into a software product line in *ABS*, thereby significantly reducing code size and increasing maintainability and understandability, which was also aided by the higher-level language features of *ABS*.

Besides simulation and state space search using the interpreter, a current line of research concentrates on the automated extraction of a verifiable model for model checking. The approach assumes finite data and a bound on the messages and translates *ABS* models to the input language for MCMAS [52], a model checker for multi agent systems. While this approach does not allow verification of large systems, it integrates the presented approach by providing new techniques for test case generation and abstraction. A publication on this topic is currently under submission.

The main objective of this study was to evaluate how *Creol* and *ABS* can be applied to complex, distributed algorithms in networks. We found the *Creol* and *ABS* languages and their tools useful in the evaluation of functional properties of the AODV algorithm and we gained insight into how complex algorithms like AODV work. We observed how small changes in the algorithm, and in the chosen arrangement, affect its behaviour so that certain properties fail. We studied these

properties in the implementation of our model, which will lead to further investigation of the reasons for its behaviour or misbehaviour, and the development of better formalisms and languages for modelling and evaluation of properties.

Systems such as *Vereofy* are more specific towards model checking using the automaton approach that is farther from real programs than *ABS* and *Creol*. On the other side, simulators, such as ns-2, ns-3, or OMNeT++, are better suited to evaluate the values of properties, such as timing or power consumption. However, these do not offer native facilities for model checking. Thus, the choice of language and simulating system is highly dependent on the question to a model and a simulation.

The choice of evaluation and simulation tools for distributed algorithms depends on the goal of the evaluation. *ABS* is a further development of *Creol* and contains most of *Creol*'s features that have shown to be useful. *ABS* has mended some of the restrictions of *Creol*. *ABS* and *Creol* belong to the languages that allow executable models that are suited for both simulation, model checking, and testing.

Acknowledgements

This research is in the context of the EU projects IST-33826 CREDO: *Modeling and analysis of evolutionary structures for distributed services*, FP7-231620 HATS: *Highly Adaptable and Trustworthy Software using Formal Models*, and FP7-PEOPLE-252184 DiVerMAS: *Distributed Systems Verification with MAS-based Model Checking*. The authors want to express their thanks to their colleagues involved in these projects for their support during this work, especially Sascha Klüppelholz, Joachim Klein, Immo Grabe, Bjarte M. Østvold, Xuedong Liang, Marcel Kyas, Martin Steffen, and in memoriam Tobias Blechmann. The authors are grateful to Trenton Schulz and the anonymous reviewers for pointing out improvements in earlier versions of this paper.

References

[1] W. Leister, J. Bjørk, R. Schlatte, and A. Griesmayer, "Verifying distributed algorithms with executable Creol models," *Proc. First International Conference on Performance, Safety and Robustness in Complex Systems and Applications (PESARO'11)*, 2011.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

[3] E. B. Johnsen, R. Hähnle, J. Schäfer, R. Schlatte, and M. Steffen, "ABS: A core language for abstract behavioral specification," in *Proc. 9th International Symposium on Formal Methods for Components and Objects (FMCO'10)*, ser. Lecture Notes in Computer Science, B. Aichernig, F. S. de Boer, and M. M. Bonsangue, Eds., vol. 6957. Springer-Verlag, 2012, pp. 142–164.

[4] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," RFC 3561 (Experimental), Jul. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3561.txt

[5] W. Leister, J. Bjørk, R. Schlatte, and A. Griesmayer, "Validation of creol models for routing algorithms in wireless sensor networks," Norsk Regnesentral, Oslo, Norway, Report 1024, 2010.

[6] K. Pohl, G. Böckle, and F. van der Linden, *Software Product Line Engineering: Foundations, Principles, and Techniques*. Springer-Verlag, 2005.

[7] I. Schaefer and R. Hähnle, "Formal methods in software product line engineering," *IEEE Computer*, vol. 44, no. 2, pp. 82–85, Feb. 2011.

[8] Q. Boucher, A. Classen, P. Faber, and P. Heymans, "Introducing TVL, a text-based feature modelling language," in *Proc. 4th International Workshop on Variability Modelling of Software-intensive Systems (VaMoS'10)*. University of Duisburg-Essen, Jan 2010, pp. 159–162.

[9] N. Medvidovic and R. Taylor, "A Classification and Comparison Framework for Software Architecture Description Languages," *IEEE Transactions on Software Engineering*, vol. 26, no. 1, pp. 70–93, 2000.

[10] J. E. Rumbaugh, I. Jacobson, and G. Booch, *The unified modeling language reference manual*. Addison-Wesley-Longman, 1999.

[11] C. Atkinson, J. Bayer, C. Bunse, E. Kamsties, O. Laitenberger, R. Laqua, D. Muthig, B. Paech, J. Wüst, and J. Zettel, *Component-based product line engineering with UML*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[12] H. Gomaa, *Designing Software Product Lines with UML*. Addison Wesley, 2004.

[13] T. Ziadi, L. Hélouët, and J.-M. Jézéquel, "Towards a UML profile for software product lines," in *Revised Papers 5th Workshop on Product Familiy Engineering (PFE'03)*, ser. Lecture Notes in Computer Science, vol. 3014. Springer Berlin / Heidelberg, 2004, pp. 129–139.

[14] W. Leister, "CreolE — a pragmatic extension to Creol," Norsk Regnesentral, Note DART/06/09, August 2009.

[15] E. B. Johnsen and O. Owe, "An asynchronous communication model for distributed concurrent objects," *Software and Systems Modeling*, vol. 6, no. 1, pp. 35–58, 2007.

[16] M. Kyas, *Creol Tools User Guide*, 0.0n ed., Institutt for Informatikk, Universitetet i Oslo, Postboks 1080 Blindern, 0316 Oslo, Norway, May 2009.

[17] I. Grabe, M. M. Jaghoori, B. Aichernig, T. Blechmann, F. de Boer, A. Griesmayer, E. B. Johnsen, J. Klein, S. Klüppelholz, M. Kyas, W. Leister, R. Schlatte, A. Stam, M. Steffen, S. Tschirner, X. Liang, and W. Yi, "Credo methodology. Modeling and analyzing a peer-to-peer system in Credo," in *3rd International Workshop on Harnessing Theories for Tool Support in Software (TTSS'09)*, 2009.

[18] F. S. de Boer, D. Clarke, and E. B. Johnsen, "A complete guide to the future," in *Proc. 16th European Symposium on Programming (ESOP'07)*, ser. Lecture Notes in Computer Science, R. de Nicola, Ed., vol. 4421. Springer-Verlag, Mar. 2007, pp. 316–330.

[19] M. Clavel, F. Durán, S. Eker, P. Lincoln, N. Martí-Oliet, J. Meseguer, and J. F. Quesada, "Maude: Specification and programming in rewriting logic," *Theoretical Computer Science*, vol. 285, no. 2, pp. 187–243, 2002.

[20] D. Clarke, N. Diakov, R. Hähnle, E. B. Johnsen, I. Schaefer, J. Schäfer, R. Schlatte, and P. Y. H. Wong, "Modeling spatial and temporal variability with the HATS abstract behavioral modeling language," in *Proc. 11th Intl. School on Formal Methods for the Design of Computer, Communication and Software Systems (SFM 2011)*, ser. Lecture Notes in Computer Science, M. Bernardo and V. Issarny, Eds., vol. 6659. Springer-Verlag, 2011, pp. 417–457.

[21] J. Bjørk, F. S. de Boer, E. B. Johnsen, R. Schlatte, and S. L. Tapia Tarifa, "User-defined schedulers for real-time concurrent objects," *Innovations in Systems and Software Engineering*, pp. 1–15, 2012.

[22] G. A. Agha, *ACTORS: A Model of Concurrent Computations in Distributed Systems*. Cambridge, Mass.: The MIT Press, 1986.

[23] P. Haller and M. Odersky, "Scala actors: Unifying thread-based and event-based programming," *Theoretical Computer Science*, vol. 410, no. 2–3, pp. 202–220, 2009.

[24] I. Schaefer, L. Bettini, V. Bono, F. Damiani, and N. Tanzarella, "Delta-oriented programming of software product lines," in *Proc. 14th International Conference on Software Product Lines (SPLC'10)*, ser. Lecture Notes in Computer Science, J. Bosch and J. Lee, Eds., vol. 6287. Springer, 2010, pp. 77–91.

[25] D. Batory, J. Sarvela, and A. Rauschmayer, "Scaling step-wise refinement," *IEEE Trans. Software Eng.*, vol. 30, no. 6, 2004.

[26] K. Beck, *Extreme Programming*. Addison-Wesley, 1999.

[27] J. Bjørk, F. S. de Boer, E. B. Johnsen, R. Schlatte, and S. L. Tapia Tarifa, "User-defined schedulers for real-time concurrent objects," *Innovations in Systems and Software Engineering*, 2012, to appear.

[28] I. Stojmenovic, "Simulations in wireless sensor and ad hoc networks: matching and advancing models, metrics, and solutions," *IEEE Communications Magazine*, vol. 46, no. 12, pp. 102–107, 2008.

[29] M. Musuvathi, D. Y. W. Park, A. Chou, D. R. Engler, and D. L. Dill, "CMC: a pragmatic approach to model checking real code," *ACM SIGOPS Operating Systems Review*, vol. 36, no. SI, pp. 75–88, 2002.

[30] M. M. R. Mozumdar, F. Gregoretti, L. Lavagno, L. Vanzago, and S. Olivieri, "A framework for modeling, simulation and automatic code generation of sensor network application," in *Proc. Fifth Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON'08)*. IEEE, 2008, pp. 515–522.

[31] A. Varga and R. Hornig, "An overview of the OMNeT++ simulation environment," in *Proc. 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops (Simutools'08)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, pp. 60:1–60:10.

[32] OPNET, "Application and network performance," web pages, last accessed February 6, 2011. [Online]. Available: www.opnet.com

[33] OMNeT++, "Bibliography," web pages, last accessed February 6, 2011. [Online]. Available: omnetpp.org/publications

[34] ns-2, "The Network Simulator ns-2," web pages, last accessed February 6, 2011. [Online]. Available: www.isi.edu/nsnam/ns/

[35] ns-3, "Papers," wiki pages, last accessed February 6, 2011. [Online]. Available: www.nsnam.org/wiki/index.php/Papers

[36] MathWorks, "Solutions," web pages, last accessed February 6, 2011. [Online]. Available: www.mathworks.se/solutions

[37] O. Wibling, J. Parrow, and A. N. Pears, "Automatized verification of ad hoc routing protocols," in *Proc. Formal Techniques for Networked and Distributed Systems (FORTE'04)*, ser. Lecture Notes in Computer Science, vol. 3235. Springer, 2004, pp. 343–358.

[38] ——, "Ad hoc routing protocol verification through broadcast abstraction," in *Proc. Formal Techniques for Networked and Distributed Systems (FORTE'05)*, ser. Lecture Notes in Computer Science, vol. 3731. Springer, 2005, pp. 128–142.

[39] S. Chiyangwa and M. Z. Kwiatkowska, "A timing analysis of AODV," in *Proc. Formal Methods for Open Object-Based Distributed Systems (FMOODS'05)*, ser. Lecture Notes in Computer Science, vol. 3535. Springer, 2005, pp. 306–321.

[40] S. Tschirner, X. Liang, and W. Yi, "Model-based validation of QoS properties of biomedical sensor networks," in *Proc. 7th ACM international conference on Embedded software (EMSOFT'08)*. New York, NY, USA: ACM, 2008, pp. 69–78.

[41] C. Baier, T. Blechmann, J. Klein, and S. Klüppelholz, "Formal verification of components and connectors," in *Proc. 7th Software Technologies Concertation on Formal Methods for Components and Objects (FMCO'08)*, ser. Lecture Notes in Computer Science, vol. 5751. Springer-Verlag, 2009, pp. 82–101.

[42] S. Klüppelholz, "Verification of branching-time and alternating-time properties for exogenous coordination models," Ph.D. dissertation, Technische Universität Dresden, Germany, 2012.

[43] C. Baier, T. Blechmann, J. Klein, S. Klüppelholz, and W. Leister, "Design and verification of systems with exogeneous coordination using Vereofy," in *Proc. 4th Intl. Symp. on Leveraging Applications of Formal Methods, Verification and Validation (ISoLA 2010), Part II*, ser. Lecture Notes in Computer Science, vol. 6416. Springer-Verlag, 2010, pp. 97–111.

[44] P. C. Ölveczky and J. Meseguer, "Semantics and pragmatics of Real-Time Maude," *Higher-Order and Sym-*

bolic Computation, vol. 20, no. 1-2, pp. 161–196, 2007.

[45] P. C. Ölveczky and S. Thorvaldsen, "Formal modeling and analysis of the OGDC wireless sensor network algorithm in Real-Time Maude," in *Formal Methods for Open Object-Based Distributed Systems (FMOODS'07)*, ser. Lecture Notes in Computer Science, vol. 4468. Springer, 2007, pp. 122–140.

[46] E. B. Johnsen, O. Owe, J. Bjørk, and M. Kyas, "An object-oriented component model for heterogeneous nets," in *Proc. 6th International Symposium on Formal Methods for Components and Objects (FMCO'07)*, ser. Lecture Notes in Computer Science, vol. 5382. Springer, 2007, pp. 257–279.

[47] W. Leister, X. Liang, S. Klüppelholz, J. Klein, O. Owe, F. Kazemeyni, J. Bjørk, and B. M. Østvold, "Modelling of biomedical sensor networks using the Creol tools," Norsk Regnesentral, Oslo, Norway, Report 1022, 2009.

[48] J. Meseguer and G. Rosu, "The rewriting logic semantics project: A progress report," in *FCT*, ser. Lecture Notes in Computer Science, O. Owe, M. Steffen, and J. A. Telle, Eds., vol. 6914. Springer-Verlag, 2011, pp. 1–37.

[49] E. Albert, S. Genaim, M. Gómez-Zamalloa, E. B. Johnsen, R. Schlatte, and S. L. T. Tarifa, "Simulating concurrent behaviors with worst-case cost bounds," in *Proc. 17th International Symposium on Formal Methods (FM'11)*, ser. Lecture Notes in Computer Science, M. Butler and W. Schulte, Eds., vol. 6664. Springer, 2011, pp. 353–368.

[50] E. B. Johnsen, R. Schlatte, and S. L. Tapia Tarifa, "A formal model of object mobility in resource-restricted deployment scenarios," in *Proc. 8th International Symposium on Formal Aspects of Component Software (FACS 2011)*, ser. Lecture Notes in Computer Science, F. Arbab and P. Ölveczky, Eds. Springer-Verlag, 2012, to appear.

[51] A. Stam *et al.*, "Project deliverable D6.4: Validation," EU project IST-33826 CREDO: Modeling and Analysis of Evolutionary Structures for Distributed Services, Tech. Rep., 2009, accessed June 18, 2012. [Online]. Available: http://projects.cwi.nl/credo/publications/Deliverables/DeliverableD6-4.pdf

[52] A. Lomuscio, H. Qu, and F. Raimondi, "MCMAS: A model checker for the verification of multi-agent systems," in *Proc. 21st International Conference on Computer Aided Verification, CAV'09*, ser. Lecture Notes in Computer Science, A. Bouajjani and O. Maler, Eds., vol. 5643. Springer, 2009, pp. 682–688.

The Bit Error Rate for Complex SSC/MRC Combiner at Two Time Instants in the Presence of Hoyt Fading

Dragana Krstić
Department of Telecommunications,
Faculty of Electronic Engineering,
University of Niš
Aleksandra Medvedeva 14
Niš, Serbia
dragana.krstic@elfak.ni.ac.rs

Goran Stamenović
Tigar, Pirot, Serbia
goran.stamenovic@tigar.com

Petar Nikolić
Tigartyres, Pirot, Serbia
nikpetar@gmail.com

Mihajlo Stefanović
Faculty of Electronic Engineering
University of Niš
Aleksandra Medvedeva 14
Niš, Serbia
mihajlo.stefanovic@elfak.ni.ac.rs

Abstract **—The expressions for probability density function (PDF) of the Switch and Stay Combiner (SSC) output signal to noise ratio (SNR) at one time instant and the joint probability density function of the SSC combiner output signal to noise ratio at two time instants in the presence of Hoyt fading are determined in this paper Then, these expressions are used for calculation of the bit error rate for complex SSC/MRC (Switch and Stay Combining/Maximal Ratio Combining) combiner versus some parameter values. The results are shown graphically in some figures and the analysis of the parameters influence and different types of combiners is given.**

Keywords - Probability Density Function; Joint Probability Density Function; Bit Error Rate, Hoyt Fading; SSC/MRC Combiner

I. INTRODUCTION

The joint probability density function of the SSC combiner output signal at two time instants in the presence of Hoyt fading is done in [1] and based on it the bit error rate for complex SSC/MRC combiner will be calculated in this paper.

The signal propagation through wireless communications channels has received a great deal of research interest [2]-[4]. The random fluctuations of the signal envelope and phase in a radio channel are caused with two propagation phenomena: multipath scattering (fast fading) and shadowing (slow fading).

The multipath fading is modeled by several distributions such as: Rayleigh, Rice, Nakagami-*m* and Weibull. The Hoyt (Nakagami-*q*) distribution model has recently received increased attention in modeling fading channels. This fading model provides a very accurate fit to experimental channel measurements in a various communication applications, such as mobile satellite propagation channels [5]. It spans the range of the fading figure from the one-sided Gaussian to the Rayleigh distribution [6]. Similarly, the Hoyt distribution can be considered as an accurate fading model for satellite links with strong ionospheric scintillation [7]. Recently, in [8], an ergodic capacity analysis is presented, and in [9] the information outage probability of orthogonal space-time block code (OSTBC) over Hoyt fading channels has been studied. Also in [10] this model has been used in outage analysis of cellular mobile radio systems, while in [11] a capacity analysis of Hoyt fading is provided.

In wireless communication systems, various techniques for reducing fading effect and influence of shadow effects are used. Such techniques are diversity reception, dynamic channel allocation and power control. Upgrading transmission reliability and increasing channel capacity without increasing transmission power and bandwidth is the main goal of diversity techniques.

The space diversity combining techniques, based on using multiple antennas at the reception, are very efficient methods used for improving system's quality of service and ensures efficient solution for reduction of signal level fluctuations in channels with fading. Multiple received copies of signal could be combined on different ways among which the most popular are: maximal ratio combining (MRC), equal gain combining (EGC), and generalized selection combining (GSC) [2]-[4]. Their complexity of implementation is relatively high since they require a separate channel for each diversity branch.

Between the simpler diversity combining schemes, the most popular are selection combining (SC) and switch and stay combining (SSC). SC and SSC types of diversity systems process only one of the diversity branches, so they are less complicated. With SSC combining the switching of the receiver between the two receiving antennas is based on a comparison of the instantaneous SNR of the connected antenna with a predetermined threshold what results in a reduction of complexity with regard to SC combining. The simultaneous and continuous monitoring of the SNRs at

both branches is no longer necessary. The price of this simplification is some loss in performances.

Namely, in SSC combiner a particular antenna is selected until its quality drops below a predetermined threshold. When this happens, the receiver switches to another antenna and stays with it for the next time interval, regardless of whether or not the channel quality of that antenna is above or below the predetermined threshold. The consideration of SSC systems in the literature has been restricted to low-complexity mobile units where the number of diversity antennas is typically limited to two [12], [13].

II. Related Work

The use of SSC combiner with great number of branches can minimize the bit error rate (BER) [14]. Dual SSC combiner is considered because the gain is the greatest when dual SSC combiner is used instead of one-channel system. The improvement becomes less with enlarging of the branch numbers [14]. The ratio of price and complexity is the best for dual branch system. Because of that it is more economic using SSC combiner with two inputs.

The probability density function (PDF) of the SSC combiner output signal at one time instant and the joint probability density function of the SSC combiner output signal at two time instants in the presence of Rayleigh, Nakagami-m, Weibull, log-normal and Hoyt fading are determined in [15] - [18] and in [1], respectively.

The authors showed in [19]-[22], based on results obtained in [15]-[18], that the error probability and the outage probability are significantly reduced if the decision is performed in two time instants. The analysis of the complex SSC/SC combiner outage probability at two time instants in the presence of Rayleigh and log-normal fading are done in [19] and [20] and the bit error rate for complex SSC/MRC combiner at two time instants in the presence of log-normal and Rayleigh fading in [21] and [22], respectively.

Based on the expressions for the PDF of the SSC combiner output SNR at one time instant and the joint PDF of the SSC combiner output SNR at two time instants in the presence of Hoyt fading obtained in [1], the bit error rate for complex SSC/MRC combiner at two time instants in the presence of Hoyt fading will be given in this paper.

Because it is shown earlier that the better system performances are obtained by decision in two time instants for other fading influences, the motivation for this work is determination of system performances in the presence of Hoyt fading since Hoyt fading has increasing importance in the study of telecommunications systems now days. This investigation could be useful and important for designers and scientists who deal with the decision based on multiple samples and to those who study the impact of different types of fading on system performances.

The remainder of this work is organized in the following way: Section III introduces the model of dual SSC combiner and determines PDF of the SSC combiner output SNR at one time instant. In Section IV, the joint PDF of the SSC combiner output SNR at two time instants is calculated. Subsequently, in fifth Section the bit error rate calculation for complex Switch and Stay Combining/Maximal Ratio Combining (SSC/MRC) combiner is calculated and then in the sixth chapter the numerical results are presented graphically. Final part of this paper is conclusion with an analysis of the obtained results.

III. System Model and Performances at One Time Instant

The system model is presented in Fig. 1. The combiner input signals to noise ratios are γ_1 and γ_2, with γ being the combiner output signal to noise ratio.

Dual SSC combiner works in the following manner: the probability of the event that combiner first examines the signal from first input is P_1, and for the second input to be examined first is P_2. If the combiner tests first the signal from first input and if the value of the signal to noise ratio at this input is greater than the threshold, γ_T, SSC combiner leads this signal to the decision circuit. If the value of the signal to noise ratio from the first input is less then the threshold γ_T, SSC combiner directs the signal from another input to the decision circuit, regardless if it is above or below the predetermined threshold. If the SSC combiner first examines the signal from second input his working algorithm is similar.

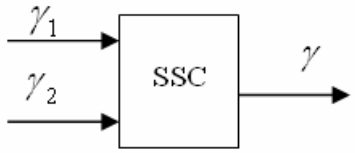


Figure 1. Model of dual SSC combiner with two inputs

The expression for PDF of the combiner output signal to noise ratio will be defined first for the case that input signal to noise ratio is less than the threshold, $\gamma < \gamma_T$. Based on the work algorithm of the SSC combiner in this case, PDF is equal:

$$p_\gamma(\gamma) = P_1 \cdot F_{\gamma_1}(\gamma_T) \cdot p_{\gamma_2}(\gamma) + P_2 \cdot F_{\gamma_2}(\gamma_T) \cdot p_{\gamma_1}(\gamma) \quad (1)$$

In the case $\gamma \geq \gamma_T$ the expression for PDF of the signal to noise ratio at the combiner output is :

$$p_\gamma(\gamma) = P_1 \cdot p_{\gamma_1}(\gamma) + P_1 \cdot F_{\gamma_1}(\gamma_T) \cdot p_{\gamma_2}(\gamma) + \\ + P_2 \cdot p_{\gamma_2}(\gamma) + P_2 \cdot F_{\gamma_2}(\gamma_T) \cdot p_{\gamma_1}(\gamma) \quad (2)$$

where γ_T is the decision threshold. The cumulative probability densities (CDFs) are given by [4]:

$$F_{\gamma_i}(\gamma_T) = \int_0^{\gamma_T} p_{\gamma_i}(x)dx, \qquad i = 1,2 \quad (3)$$

The probabilities P_1 and P_2 are [4]:

$$P_1 = \frac{F_{\gamma_2}(\gamma_T)}{F_{\gamma_1}(\gamma_T) + F_{\gamma_2}(\gamma_T)} \quad (4)$$

$$P_2 = \frac{F_{\gamma_1}(\gamma_T)}{F_{\gamma_1}(\gamma_T) + F_{\gamma_2}(\gamma_T)} \quad (5)$$

The PDFs of the SNRs at the combiner input, γ_1 and γ_2, in the presence of Hoyt fading, are [4]:

$$p_{\gamma_1}(\gamma_1) = \frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)$$

$$\gamma_1 \geq 0 \quad (6)$$

$$p_{\gamma_2}(\gamma_2) = \frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)$$

$$\gamma_2 \geq 0 \quad (7)$$

where q_i are Nakagami-q fading parameters, which range from 0 to 1 and $\overline{\gamma}_i$ are average SNRs for input channels

The CDFs of the SNRs at the combiner input in the presence of Hoyt fading, after putting of the expressions (6), (7), into (3), are given by:

$$F_{r_1}(\gamma_T) = \int_0^{\gamma_T}\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2 x}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)x}{4q_1^{\,2}\overline{\gamma}_1}\right)dx =$$

$$= \frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}}, \frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right) \quad (8)$$

$$F_{r_2}(\gamma_T) = \int_0^{\gamma_T}\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2 x}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)x}{4q_2^{\,2}\overline{\gamma}_2}\right)dx =$$

$$= \frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}}, \frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right) \quad (9)$$

where $I_e(k, x)$ is Rice's I_e function [23].

After putting of the expressions (8) and (9) into (4) and (5), the probabilities P_1 and P_2 are:

$$P_1 = \frac{\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}}, \frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)}{\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}}, \frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right) + \frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}}, \frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)} \quad (10)$$

$$P_2 = \frac{\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}}, \frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)}{\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}}, \frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right) + \frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}}, \frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)} \quad (11)$$

After putting of the expressions (4)-(7), (8) and (9) into (1), the PDF of the combiner output SNR, γ, for $\gamma < \gamma_T$, is:

$$p_\gamma(\gamma) = P_1\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}}, \frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}}, \frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)\cdot$$

$$\cdot\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right) \quad (12)$$

After putting of the expressions (4)-(7), (10) and (11) into (2), PDF of the combiner output SNR γ, for $\gamma \geq \gamma_T$ is:

$$p_\gamma(\gamma) = P_1\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)+$$

$$+P_1\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}}, \frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}}, \frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)\cdot$$

$$\cdot\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right) \quad (13)$$

The obtained expressions for the PDF of the output signal to noise ratio after diversity combining can be used to study the moments, the amount of fading, the outage probability and the average bit error rate of proposed system.

IV. SYSTEM PERFORMANCES AT TWO TIME INSTANTS

The model of dual SSC combiner at two time instants considered in this Section is shown in Fig. 2.

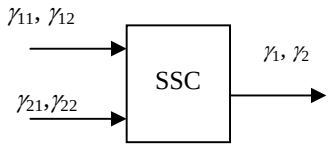


Figure 2. Model of the SSC combiner with two inputs at two time instants

The input SNRs are γ_{11} and γ_{21} at the first time instant and they are γ_{12} and γ_{22} at the the second time moment. The output SNRs are γ_1 and γ_2.

The indexes for the input SNRs are: first index is the ordinal branch number and the other signs time instant observed. For the output SNRs, the index represents the time instant observed.

The joint PDF of uncorrelated input signals, with Hoyt distribution and the same parameters, is [4]:

$$p_{\gamma_1\gamma_2}(\gamma_1,\gamma_2)=\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right) \qquad (14)$$

Modified Bessel function of the first kind is defined by [24]:

$$I_m(x)=\left(\frac{x}{2}\right)^m\sum_{k=0}^{\infty}\frac{\left(\frac{x^2}{4}\right)^k}{k!\Gamma(m+k+1)} \qquad (15)$$

Four different cases are observed. The first case is: $\gamma_1<\gamma_T$ and $\gamma_2<\gamma_T$. In this case all signal to noise ratios at the input are below γ_T, i.e.: $\gamma_{11}<\gamma_T$, $\gamma_{12}<\gamma_T$, $\gamma_{21}<\gamma_T$, and $\gamma_{22}<\gamma_T$.

Let the combiner first treat the signal r_{11}. Because $\gamma_{11}<\gamma_T$, therefore that $\gamma_1=\gamma_{21}$, and since $\gamma_{22}<\gamma_T$ it is $\gamma_2=\gamma_{12}$. The probability of this situation is P_1.

When SSC combiner first treats the signal r_{21}, then it is $\gamma_1=\gamma_{11}$, because $\gamma_{21}<\gamma_T$. After $\gamma_{12}<\gamma_T$, then it is $\gamma_2=\gamma_{22}$. The probability of this situation is P_2. After previous, the joint PDF of the combiner output SNRs at two time instants is obtained using expression (14), for $\gamma_1<\gamma_T$ and $\gamma_2<\gamma_T$:

$$p_{\gamma_1\gamma_2}(\gamma_1,\gamma_2)=P_1\cdot\int_0^{\gamma_T}p_{\gamma_{11}\gamma_{12}}(\gamma_{11},\gamma_2)d\gamma_{11}\int_0^{\gamma_T}p_{\gamma_{22}\gamma_{21}}(\gamma_{22},\gamma_1)d\gamma_{22}+$$

$$+P_2\cdot\int_0^{\gamma_T}p_{\gamma_{21}\gamma_{22}}(\gamma_{21},\gamma_2)d\gamma_{21}\int_0^{\gamma_T}p_{\gamma_{12}\gamma_{11}}(\gamma_{12},\gamma_1)d\gamma_{12}=$$

$$=P_1\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right) \qquad (16)$$

In the similar manner the other joint PDFs can be derived. The joint PDF is, for $\gamma_1\geq\gamma_T$ and $\gamma_2<\gamma_T$:

$$p_{\gamma_1\gamma_2}(\gamma_1,\gamma_2)=P_1\cdot p_{\gamma_{22}}(\gamma_2)\int_0^{\gamma_T}p_{\gamma_{12}\gamma_{11}}(\gamma_{12},\gamma_1)d\gamma_{12}+$$

$$+P_1\cdot\int_0^{\gamma_T}p_{\gamma_{11}\gamma_{12}}(\gamma_{11},\gamma_2)d\gamma_{11}\int_0^{\gamma_T}p_{\gamma_{22}\gamma_{21}}(\gamma_{22},\gamma_1)d\gamma_{22}+$$

$$+P_2\cdot p_{\gamma_{12}}(\gamma_2)\int_0^{\gamma_T}p_{\gamma_{22}\gamma_{21}}(\gamma_{22},\gamma_1)d\gamma_{22}+P_2\cdot\int_0^{\gamma_T}p_{\gamma_{21}\gamma_{22}}(\gamma_{21},\gamma_2)d\gamma_{21}\int_0^{\gamma_T}p_{\gamma_{12}\gamma_{11}}(\gamma_{12},\gamma_1)d\gamma_{12}=$$

$$=P_1\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\cdot$$

$$\cdot\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_1\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right) \qquad (17)$$

for $\gamma_1<\gamma_T$ and $\gamma_2\geq\gamma_T$:

$$p_{\gamma_1\gamma_2}(\gamma_1,\gamma_2) = P_1 \cdot \int_0^{\gamma_T} p_{\gamma_{11}}(\gamma_{11})d\gamma_{11} \cdot p_{\gamma_{21}\gamma_{22}}(\gamma_1,\gamma_2) +$$

$$+P_1 \cdot \int_0^{\gamma_T} p_{\gamma_{11}\gamma_{12}}(\gamma_{11},\gamma_2)d\gamma_{11} \int_0^{\gamma_T} p_{\gamma_{22}\gamma_{21}}(\gamma_{22},\gamma_1)d\gamma_{22} +$$

$$+P_2 \cdot \int_0^{\gamma_T} p_{\gamma_{21}}(\gamma_{21})d\gamma_{21} \cdot p_{\gamma_{11}\gamma_{12}}(\gamma_1,\gamma_2) + P_2 \cdot \int_0^{\gamma_T} p_{\gamma_{21}\gamma_{22}}(\gamma_{21},\gamma_2)d\gamma_{21} \int_0^{\gamma_T} p_{\gamma_{12}\gamma_{11}}(\gamma_{12},\gamma_1)d\gamma_{12} =$$

$$= P_1 \frac{2q_1}{1+q_1^2} I_e\left(\frac{1-q_1^2}{1+q_1^2}, \frac{\left(1+q_1^2\right)^2 \gamma_T}{4q_1^2\overline{\gamma}_1}\right) \cdot$$

$$\cdot \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_1}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_1}{4q_2^2\overline{\gamma}_2}\right) \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_2}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_2}{4q_2^2\overline{\gamma}_2}\right) +$$

$$+P_2 \frac{2q_2}{1+q_2^2} I_e\left(\frac{1-q_2^2}{1+q_2^2}, \frac{\left(1+q_2^2\right)^2 \gamma_T}{4q_2^2\overline{\gamma}_2}\right) \cdot$$

$$\cdot \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_1}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_1}{4q_1^2\overline{\gamma}_1}\right) \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_2}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_2}{4q_1^2\overline{\gamma}_1}\right) +$$

$$+P_1 \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_2}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_2}{4q_1^2\overline{\gamma}_1}\right) \frac{2q_1}{1+q_1^2} I_e\left(\frac{1-q_1^2}{1+q_1^2}, \frac{\left(1+q_1^2\right)^2 \gamma_T}{4q_1^2\overline{\gamma}_1}\right) \cdot$$

$$\cdot \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_1}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_1}{4q_2^2\overline{\gamma}_2}\right) \frac{2q_2}{1+q_2^2} I_e\left(\frac{1-q_2^2}{1+q_2^2}, \frac{\left(1+q_2^2\right)^2 \gamma_T}{4q_2^2\overline{\gamma}_2}\right) +$$

$$+P_2 \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_1}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_1}{4q_1^2\overline{\gamma}_1}\right) \frac{2q_1}{1+q_1^2} I_e\left(\frac{1-q_1^2}{1+q_1^2}, \frac{\left(1+q_1^2\right)^2 \gamma_T}{4q_1^2\overline{\gamma}_1}\right) \cdot$$

$$\cdot \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_2}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_2}{4q_2^2\overline{\gamma}_2}\right) \frac{2q_2}{1+q_2^2} I_e\left(\frac{1-q_2^2}{1+q_2^2}, \frac{\left(1+q_2^2\right)^2 \gamma_T}{4q_2^2\overline{\gamma}_2}\right) \tag{18}$$

for $\gamma_1 \geq \gamma_T$ and $\gamma_2 \geq \gamma_T$:

$$p_{\gamma_1\gamma_2}(\gamma_1,\gamma_2) = P_1 \cdot p_{\gamma_{11}\gamma_{12}}(\gamma_1,\gamma_2) + P_1 \cdot p_{\gamma_{22}}(\gamma_2) \int_0^{\gamma_T} p_{\gamma_{12}\gamma_{11}}(\gamma_{12},\gamma_1)d\gamma_{12} +$$

$$+P_1 \cdot \int_0^{\gamma_T} p_{\gamma_{11}}(\gamma_{11})d\gamma_{11} \cdot p_{\gamma_{21}\gamma_{22}}(\gamma_1,\gamma_2) + P_1 \cdot \int_0^{\gamma_T} p_{\gamma_{11}\gamma_{12}}(\gamma_{11},\gamma_2)d\gamma_{11} \int_0^{\gamma_T} p_{\gamma_{22}\gamma_{21}}(\gamma_{22},\gamma_1)d\gamma_{22} +$$

$$+ P_2 \cdot p_{\gamma_{21}\gamma_{22}}(\gamma_1,\gamma_2) + P_2 \cdot p_{\gamma_{12}}(\gamma_2) \int_0^{r_T} p_{\gamma_{22}\gamma_{21}}(\gamma_{22},\gamma_1)d\gamma_{22} +$$

$$+P_2 \cdot \int_0^{\gamma_T} p_{\gamma_{21}}(\gamma_{21})d\gamma_{21} \cdot p_{\gamma_{11}\gamma_{12}}(\gamma_1,\gamma_2) + P_2 \cdot \int_0^{\gamma_T} p_{\gamma_{21}\gamma_{22}}(\gamma_{21},\gamma_2)d\gamma_{21} \int_0^{\gamma_T} p_{\gamma_{12}\gamma_{11}}(\gamma_{12},\gamma_1)d\gamma_{12} =$$

$$= P_1 \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_1}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_1}{4q_1^2\overline{\gamma}_1}\right) \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \cdot$$

$$\cdot \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_2}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_2}{4q_1^2\overline{\gamma}_1}\right) +$$

$$+P_1 \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_2}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_2}{4q_2^2\overline{\gamma}_2}\right) \cdot$$

$$\cdot \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_1}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_1}{4q_1^2\overline{\gamma}_1}\right) \frac{2q_1}{1+q_1^2} I_e\left(\frac{1-q_1^2}{1+q_1^2}, \frac{\left(1+q_1^2\right)^2 \gamma_T}{4q_1^2\overline{\gamma}_1}\right) +$$

$$+P_1 \frac{2q_1}{1+q_1^2} I_e\left(\frac{1-q_1^2}{1+q_1^2}, \frac{\left(1+q_1^2\right)^2 \gamma_T}{4q_1^2\overline{\gamma}_1}\right) \cdot$$

$$\cdot \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_1}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_1}{4q_2^2\overline{\gamma}_2}\right) \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \cdot$$

$$\cdot \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_2}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_2}{4q_2^2\overline{\gamma}_2}\right) +$$

$$+P_1 \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_2}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_2}{4q_1^2\overline{\gamma}_1}\right) \frac{2q_1}{1+q_1^2} I_e\left(\frac{1-q_1^2}{1+q_1^2}, \frac{\left(1+q_1^2\right)^2 \gamma_T}{4q_1^2\overline{\gamma}_1}\right) \cdot$$

$$\cdot \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_1}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_1}{4q_2^2\overline{\gamma}_2}\right) \frac{2q_2}{1+q_2^2} I_e\left(\frac{1-q_2^2}{1+q_2^2}, \frac{\left(1+q_2^2\right)^2 \gamma_T}{4q_2^2\overline{\gamma}_2}\right) +$$

$$+P_2 \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_1}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_1}{4q_2^2\overline{\gamma}_2}\right) \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \cdot$$

$$\cdot \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_2}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_2}{4q_2^2\overline{\gamma}_2}\right) +$$

$$+P_2 \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_2}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_2}{4q_1^2\overline{\gamma}_1}\right) \cdot$$

$$\cdot \frac{\left(1+q_2^2\right)}{2q_2\overline{\gamma}_2} \exp\left(-\frac{\left(1+q_2^2\right)^2 \gamma_1}{4q_2^2\overline{\gamma}_2}\right) I_0\left(\frac{\left(1-q_2^4\right)\gamma_1}{4q_2^2\overline{\gamma}_2}\right) \frac{2q_2}{1+q_2^2} I_e\left(\frac{1-q_2^2}{1+q_2^2}, \frac{\left(1+q_2^2\right)^2 \gamma_T}{4q_2^2\overline{\gamma}_2}\right) +$$

$$+P_2 \frac{2q_2}{1+q_2^2} I_e\left(\frac{1-q_2^2}{1+q_2^2}, \frac{\left(1+q_2^2\right)^2 \gamma_T}{4q_2^2\overline{\gamma}_2}\right) \cdot$$

$$\cdot \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \exp\left(-\frac{\left(1+q_1^2\right)^2 \gamma_1}{4q_1^2\overline{\gamma}_1}\right) I_0\left(\frac{\left(1-q_1^4\right)\gamma_1}{4q_1^2\overline{\gamma}_1}\right) \frac{\left(1+q_1^2\right)}{2q_1\overline{\gamma}_1} \cdot$$

$$\cdot\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right) \tag{19}$$

V. BIT ERROR RATE FOR SSC/MRC COMBINER

The model of the SSC/MRC combiner with two inputs considered in this paper is shown in Fig. 3. We consider the SSC/MRC combiner with two branches at two time instants. The output signals at SSC part are γ_1 and γ_2 and they become the inputs at MRC combiner and the overall output signal is γ. In this case, the signal at output of complex combiner includes both, the time diversity and the space diversity {1}.

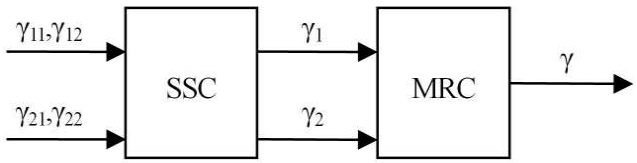


Figure 3. Model of complex dual SSC/MRC combiner at two time instants

The total conditional signal value at the output of the MRC combiner, for equally transmitted symbols of L branch MRC receiver, is given by [3]

$$\gamma=\sum_{l=1}^{L}\gamma_l \tag{20}$$

For coherent binary signals the conditional BER $P_b\left(e\middle|\{\gamma_l\}_{l=1}^{L}\right)$ is given by [4]

$$P_b\left(e\middle|\{\gamma_l\}_{l=1}^{L}\right)=Q\left(\sqrt{2g\gamma}\right) \tag{21}$$

where Q is the one-dimensional Gaussian Q-function [17]

$$Q(x)=\frac{1}{\sqrt{2\pi}}\int_x^{\infty}e^{-t^2/2}dt \tag{22}$$

Gaussian Q-function can be defined using alternative form as [4]

$$Q(x)=\frac{1}{\pi}\int_0^{\pi/2}\exp\left(-\frac{x^2}{2\sin^2\phi}\right)d\phi \tag{23}$$

Using the alternative representation of the Gaussian-Q function, the conditional BER can be expressed as

$$P_b\left(e\middle|\{\gamma_l\}_{l=1}^{L}\right)=\frac{1}{\pi}\int_0^{\pi/2}\exp\left(-\frac{g\gamma}{\sin^2\phi}\right)d\phi=\frac{1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\left(-\frac{g\gamma_l}{\sin^2\phi}\right)d\phi \tag{24}$$

The unconditional BER can be obtained by averaging the multichannel conditional BER over the joint PDF of the signals at the input of MRC combiner

$$P_b(e)=\underbrace{\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}}_{L}P_b\left(\{\gamma_l\}_{l=1}^{L}\right)\prod_{l=1}^{L}p_{\gamma_1,\gamma_2,\ldots\gamma_L}(\gamma_1,\gamma_2,\ldots\gamma_L)d\gamma_1d\gamma_2\ldots d\gamma_L \tag{25}$$

Substituting (24) in (25), $P_b(e)$ is obtained as

$$P_b(e)=\underbrace{\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}\frac{1}{\pi}\int_0^{\pi/2}\prod_{l=1}^{L}\left(-\frac{g\gamma_l}{\sin^2\phi}\right)d\phi}_{L}p_{\gamma_1,\gamma_2,\ldots\gamma_L}(\gamma_1,\gamma_2,\ldots\gamma_L)d\gamma_1d\gamma_2\ldots d\gamma_L \tag{26}$$

For dual branch MRC combiner, $P_b(e)$ is

$$\begin{aligned}P_b(e)=&\frac{1}{\pi}\int_0^{\gamma_T}\int_0^{\gamma_T}\int_0^{\pi/2}\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)p_{\gamma_1,\gamma_2}(\gamma_1,\gamma_2)d\phi d\gamma_1d\gamma_2+\\&+\frac{1}{\pi}\int_{\gamma_T}^{\infty}\int_0^{\gamma_T}\int_0^{\pi/2}\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)p_{\gamma_1,\gamma_2}(\gamma_1,\gamma_2)d\phi d\gamma_1d\gamma_2+\\&+\frac{1}{\pi}\int_0^{\gamma_T}\int_{\gamma_T}^{\infty}\int_0^{\pi/2}\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)p_{\gamma_1,\gamma_2}(\gamma_1,\gamma_2)d\phi d\gamma_1d\gamma_2+\\&+\frac{1}{\pi}\int_{\gamma_T}^{\infty}\int_{\gamma_T}^{\infty}\int_0^{\pi/2}\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)p_{\gamma_1,\gamma_2}(\gamma_1,\gamma_2)d\phi d\gamma_1d\gamma_2\end{aligned} \tag{27}$$

Substituting (1-4) in (27), $P_b(e)$ of SSC/MRC combiner can be obtained as:

$$P_b(e)=\frac{1}{\pi}\int_0^{\gamma_T}\int_0^{\gamma_T}\int_0^{\pi/2}d\gamma_1d\gamma_2d\phi\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)\cdot$$

$$\cdot P_1\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+\frac{1}{\pi}\int_{\gamma_T}^{\infty}\int_0^{\gamma_T}\int_0^{\pi/2}d\gamma_1d\gamma_2d\phi\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)\cdot$$

$$\cdot P_1 \frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\cdot$$

$$\cdot\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_1\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+\frac{1}{\pi}\int_0^{\gamma_T}\int_{\gamma_T}^{\infty}\int_0^{\pi/2}d\gamma_1 d\gamma_2 d\phi\left(-\frac{g\gamma_1}{\sin^2\phi}\right)\left(-\frac{g\gamma_2}{\sin^2\phi}\right)\cdot\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)\cdot$$

$$\cdot\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)+$$

$$+P_1\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_2}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_1}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)+$$

$$+P_2\frac{\left(1+q_1^{\,2}\right)}{2q_1\overline{\gamma}_1}\exp\left(-\frac{\left(1+q_1^{\,2}\right)^2\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)I_0\left(\frac{\left(1-q_1^{\,4}\right)\gamma_1}{4q_1^{\,2}\overline{\gamma}_1}\right)\frac{2q_1}{1+q_1^{\,2}}I_e\left(\frac{1-q_1^{\,2}}{1+q_1^{\,2}},\frac{\left(1+q_1^{\,2}\right)^2\gamma_T}{4q_1^{\,2}\overline{\gamma}_1}\right)\cdot$$

$$\cdot\frac{\left(1+q_2^{\,2}\right)}{2q_2\overline{\gamma}_2}\exp\left(-\frac{\left(1+q_2^{\,2}\right)^2\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)I_0\left(\frac{\left(1-q_2^{\,4}\right)\gamma_2}{4q_2^{\,2}\overline{\gamma}_2}\right)\frac{2q_2}{1+q_2^{\,2}}I_e\left(\frac{1-q_2^{\,2}}{1+q_2^{\,2}},\frac{\left(1+q_2^{\,2}\right)^2\gamma_T}{4q_2^{\,2}\overline{\gamma}_2}\right)\tag{28}$$

VI. NUMERICAL RESULTS

It is simple to present graphically the results, that come from equations we obtained in Sections III-V, using mathematical software, for example "Matlab", that is used in this paper. For the sake of simplicity, it has been assumed that the variances of both signals at the combiner input are equal without loosing in generalization.

The case when one time instant is observed is presented in Fig. 4. The PDF of the combiner output SNR is determined as a function of input SNR γ and the decision threshold γ_T, for three different variance values and for the same distribution parameters at receiver branches.

The Fig. 5. to 7. present the situation when two time instants are observed. The PDF is given versus input SNRs at two time instants, γ_1 and γ_2, for different values of distribution parameters and decision threshold γ_T.

It is assumed that both input branches have the same channel parameters. It is adopted that γ_T is optimal threshold of the SSC decision [1]:

$$\gamma_T = E(\gamma^1), \tag{29}$$

where [25]:

$$E(\gamma^k) = \Gamma(1+k)\,{}_1F_1\left(-\frac{k-1}{2},-\frac{k}{2};1;\left(\frac{1-q^2}{1+q^2}\right)^2\right)\bar{\gamma}^k. \tag{30}$$

Substituting (30) in (29), the optimal threshold is:

$$\gamma_T = \bar{\gamma}. \tag{31}$$

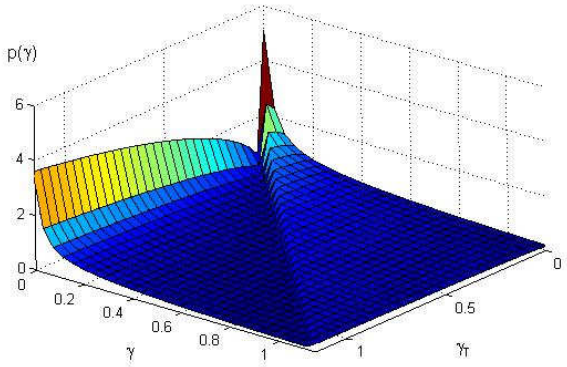


Figure 4. Probability density function of the combiner output signal to noise ratio at one time instant for $\bar{\gamma}_1 = \bar{\gamma}_2 = 1$, $q_1=q_2=0.5$

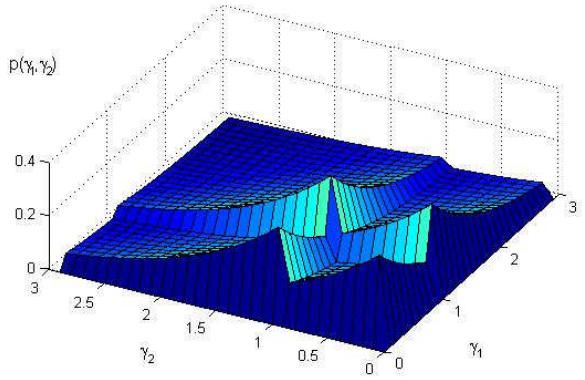


Figure 6. The probability density function of the combiner output signal to noise ratio at two time instants for $\bar{\gamma}_1 = \bar{\gamma}_2 = 1, \gamma_T = 1$, $q_1=q_2=0.9$

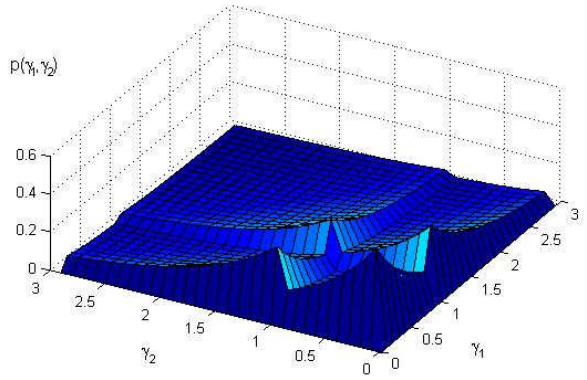


Figure 5. The probability density function of the combiner output signal to noise ratio at two time instants for $\bar{\gamma}_1 = \bar{\gamma}_2 = 1, \gamma_T = 1$, $q_1=q_2=0.5$

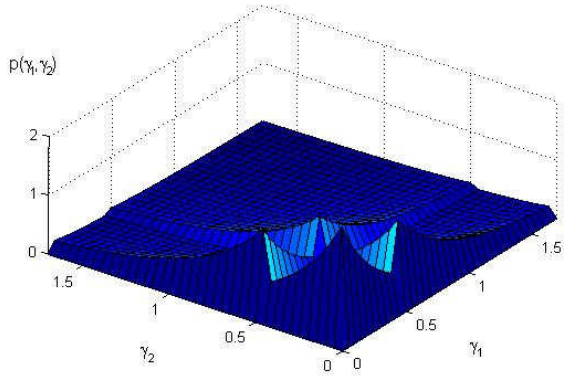


Figure 7. The probability density function of the combiner output signal to noise ratio at two time instants for $\bar{\gamma}_1 = \bar{\gamma}_2 = 0.5, \gamma_T = 0.5$, $q_1=q_2=0.5$

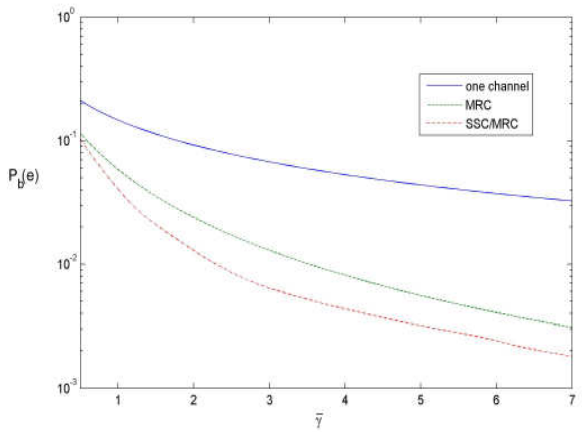


Figure 8. Bit error rate for different types of combiners versus parameter $\bar{\gamma}$ for *q=0.9*

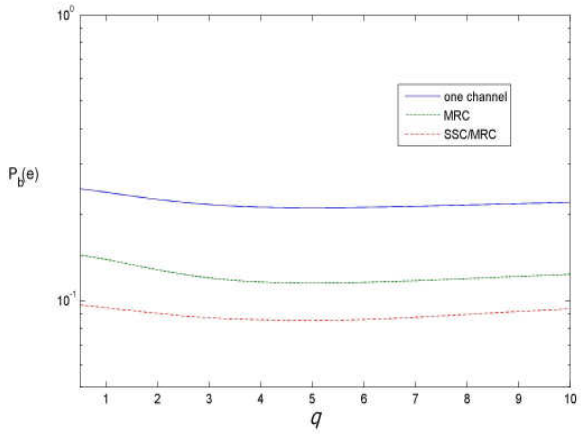


Figure 9. Bit error rate for different types of combiners versus parameter q for $\bar{\gamma} = 0.5$

The families of curves for BER for one channel receiver and for MRC combiner at one time instant and SSC/MRC combiner at two time instants, for uncorrelated case, are shown in Figs. 8. and 9. versus different distribution parameters.

It can be observed that SSC/MRC combiner has significantly better performances for uncorrelated case then MRC combiner at one time instant. It is evident that using this complex SSC/MRC combiner results in better system performances.

VII. Conclusion and Future Work

The SSC and MRC are simple and frequently used techniques for combining signals in diversity systems. The system performances deciding by two samples can be determined by the joint PDF of dual SSC combiner output signal at two time instants and putting them as inputs of the MRC combiner. In this paper, the PDF of dual SSC/MRC combiner output signal at two time instants in the presence of Hoyt fading is determined and based on it the bit error probability is calculated.

The obtained results are shown graphically for different parameters. It is apparent that system performances can be significantly improved using the sampling at two time instants and characteristics of complex SSC/MRC combiner comparing with classical SSC and MRC combiners.

In the future works, the other important useful performance measure for diversity systems like outage probability, amount of fading and second order parameters can be determined.

This fact shows that the results obtained in this paper are very significant in the designing and application of diversity receivers.

Acknowledgment

This work has been funded by the Serbian Ministry for Science under the projects TR33035, TR32052 and III 44006.

References

[1] D. Krstić, P. Nikolić, G. Stamenović, and M. Stefanović, “The joint probability density function of the SSC Combiner Output Signal at two Time Instants in the Presence of Hoyt Fading”, The Seventh International Conference on Wireless and Mobile Communications, ICWMC 2011, Luxembourg, 19-24. June 2011, IARIA XPS Press, ISBN: 978-1-61208-140-3 .

[2] W.C.Y. Lee, Mobile communications engineering, Mc-Graw-Hill, New York, 0-7803-7005-8/01, 1992.

[3] G. L. Stuber, Principles of Mobile Communications. Norwell, MA: Kluwer Academic Publishers, 1996.

[4] M. K. Simon, M. S. Alouni, Digital Communication over Fading Channels, Second Edition, Wiley-Interscience, A John Wiley&Sons, Inc., Publications, New Jersey, 2005.

[5] C-X. Wang, N. Youssef, and M. Patzold, "Level-crossing rate and average duration of fades of deterministic simulation models for Nakagami-Hoyt fading channels", 5th International Symposium on Wireless Personal Multimedia Communications, vol.1, pp. 272-276, Oct. 2002.

[6] G. Fraidenraich, J. Filho, and M. Yacoub, "Second-order statistics of Maximal-Ratio and Equal-Gain Combining in Hoyt fading", IEEE Commun. Lett, vol. 9, No 1, pp. 19-21, Jan. 2005.

[7] B. Chytil, “The distribution of amplitude scintillation and the conversion of scintillation indices,” J. Atmos. Terr. Phys., vol. 29, pp. 1175-1177, Sept. 1967.

[8] S. Khatalin and J. P. Fonseka, “On the channel capacity in Rician and Hoyt fading environments with MRC diversity,” IEEE Trans. Veh. Technol., vol. 55, no. 1, pp. 137–141, Jan. 2006.

[9] G. A. Ropokis, A. A. Rontogiannis, and P. T. Mathiopoulos, “Information outage probability of orthogonal space-time block codes over Hoyt fading channels,” in International Workshop on Satellite and Space Communications, vol. 2, pp. 135–139, Sept. 2006.

[10] A. Anamalai, C. Tellambura, and V. K. Bhargava, “Simple and accurate methods for the outage analysis in cellular mobile radio systems-A unified approach,” IEEE Trans. Commun., vol. 49, no. 2, pp. 303–316, Feb. 2001.

[11] J. Cheng and T. Berger, “Capacity of Nakagami-q (Hoyt) fading channels with channel side information,” in Proc. IEEE International Conference on Communication Technology, vol. 55, pp. 1915–1918, Apr. 2003.

[12] M. S. Alouini, and M. K. Simon, ”Postdetection Switched Combining- A simple Diversity Scheme with Improved BER

Performance", IEEE Trans. on Commun., vol. 51, No 9, pp.1591-1602, Sept 2003.

[13] Y. C. Ko, M. S. Alouini, and M. K. Simon, "Analysis and optimization of switched diversity systems", IEEE Trans. Veh. Technol., vol. 49, Issue 5, pp.1813-1831, Sept. 2000.

[14] B. Nikolic and G. Djordjevic, "Performance of SC and SSC Receivers in Hoyt Channel in the Presence of Imperfect Reference Signal Extraction", 9th International Conference on Applied Electromagnetics, Niš, Serbia, Aug. 31 – Sept. 02, 2009.

[15] M. Č. Stefanović., D. S. Krstić, M. Petrović, and Dj. Bandjur, "The Joint Probability Density Function of the SSS Combiner Output Signal at Two Time Instants in the Presence of Rayleigh Fading", The Second European Conference on Antennas and Propagation, EuCAP 2007, Edinburgh, UK, 11-16 Nov. 2007.

[16] D. Krstić, P. Nikolić., M. Matović, A. Matović, and M. Stefanović, "The Joint Probability Density Function of the SSC Combiner Output Signal in the Presence of Nakagami-m fading" The Fourth International Conference on Wireless and Mobile Communications, ICWMC 2008, Athens/Vouliagmeni, Greece, pp. 409-416, July 27-Aug. 1, 2008, IARIA XPS Press, ISBN: 978-0-7695-3274-5.

[17] P. B. Nikolic, M. C. Stefanović, D. S. Krstić, and G. Lj. Stamenović, "The Joint Probability Density Function of the SSS Combiner Output in the Presence of Weibull Fading", XV International Symposium on Theoretical Electrical Engineering, ISTET'09, Lübeck, Germany, pp. 20-24, 22 – 24 June 2009.

[18] D. S. Krstić, P. B. Nikolić, F. Destović, and M. C. Stefanović, "The Joint Probability Density Function of the SSC Combiner Output Signal in the Presence of Log-Normal Fading in two Time Instants", Electronics and Electrical Engineering, Kaunas: Technologija ISSN 1392-1215, No. 3 (109), pp. 11-16, 2011.

[19] P. Nikolić, D. Krstić, M. Milić, and M. Stefanović, "Performance Analysis of SSC/SC Combiner at Two Time Instants in The Presence of Rayleigh Fading", Frequenz. Volume 65, Issue 11-12, pp. 319–325, Nov. 2011.

[20] M. Stefanović, P. Nikolić, D. Krstić, and V. Doljak, "Outage probability of the SSC/SC combiner at two time instants in the presence of lognormal fading", Przeglad Elektrotechniczny (Electrical Review), ISSN 0033-2097, R. 88 NR 3a/2012, pp.237-240, March 2012.

[21] D. Krstić, P. Nikolić, G.Stamenović, and M. Stefanović, "Bit error rate for SSC/MRC Combiner at Two Time Instants in The Presence of log-normal Fading", Facta Universitatis. Series Automatic Control and Robotics, ISSN 1820-6417, Vol.10, No 1, 2011, pp. 83–95, UDC 621.396.94 621.395.38 519.724

[22] D. Krstić, M. Stefanović, P. Nikolić, N. Sekulović, and V. Doljak, "The Bit Error Rate for Complex SSC/MRC Combiner in the Presence of Rayleigh Fading", to be at The third International Conference on Information and Communication Systems (ICICS 2012), Irbid 22110, Jordan, Apr. 3- 5, 2012.

[23] B. T. Tan, T. T. Tjhung, C. H. Teo, and P. Y. Leong, "Series representations for Rice's Ie function," IEEE Trans. on Communications, vol. 32, no. 12, pp. 1342-1344, Dec. 1996.

[24] I. S. Gradshteyn and I. M. Ryzhik, "Table of Integrals, Series, and Products", 5th ed. San Diego, CA: Academic Press, 1994.

[25] M. Nakagami, "The m-distribution- A general formula of intensity distribution of rapid fading," in Statistical Methods in Radio Wave Propagation, Oxford, U.K.: Pergamon Press, pp. 3–36, 1960.

Impact of Scrambling-Exact Channel Estimation on HSDPA MIMO

Wolfgang Aichmann
Nokia Siemens Networks
D-89081 Ulm, Germany
Email: wolfgang.aichmann@nsn.com

***Abstract*—Multiple-input Multiple-output (MIMO) antenna systems require a separate pilot signal for each transmitting antenna for enabling the receiver to measure all physical channels separately. As scrambling operates with pseudorandom sequences the auto-correlation is quasi-orthogonal with small deviations from exact orthogonality. In this paper the impact of these deviations is investigated, which is important especially for estimating the channel impulse response. Therefore, an analytical approach beyond simple correlation is proposed, taking into account the full scrambling sequence. Link level simulations show that the performance of high-speed downlink packet access (HSDPA) MIMO systems is significantly improved, especially for very high data rates when dual stream transmission is used in combination with cancellation of the interference between the data streams.**

***Keywords*-W-CDMA; HSDPA; MIMO; channel estimation; scrambling; crosstalk; interference cancellation;**

I. INTRODUCTION

Channel estimation plays an important role in wideband code division multiple access (W-CDMA) networks, especially for very high data rates as they are achieved with MIMO [1], [2]. It has therefore been topic of different investigations, which go beyond the conventional correlation approach. In [3], a novel type of linear minimum mean square equalizer (LMMSE) was presented, which is able to properly take into account all types of interference without being too complex for implementation in real systems, whereas [4] and [5] investigate smoothening the primary channel estimates by appropriate filtering and cancelling the inter-antenna interference, respectively. An adaptive channel estimator was proposed in [6] for system utilizing space time transmit diversity.

In order to benefit from diversity of transmitted signals achieved either by separating the transmitting antennas spatially or using different polarization directions, it is necessary to measure all physical channels from each transmitting to each receiving antenna, and transmission of more than one data stream in MIMO systems is not possible at all without resolving these channels individually (see, e.g., [7]). Therefore, each transmitting antenna has to be fed with a separate pilot sequence. For HSDPA systems [8], there are defined primary and secondary common pilot channels, which are orthogonal either by using orthogonal symbol patterns or by applying orthogonal spreading codes. In any case, the coded pilot chip level sequences are scrambled.

As scrambling is a quasi-orthogonal but not an exact orthogonal operation the channel information derived by the receiver remains distorted even under conditions being ideal beyond these distortions, i.e., absence of receiver noise and additive white Gaussian noise (AWGN) channels. In [9], an iterative approach was proposed to approximately handle this phenomenon. Because equalization of the incoming signal is cancelling the effect of multi-path propagation to a considerable extent this effect does not play a role for data detection itself. However, channel estimation, which is needed for measuring the different propagation paths has to work on un-equalized incoming data and thus is impacted significantly.

As long as only one pilot channel is present, the impact is usually negligible: Although a certain amount of energy is scattered by the scrambling operation into other codes or pilot patterns, the general shape of the signal at the output of the physical channel persists as the energy reduction is more or less proportional to the available energy, i.e., slots with low pilot signal strength also have a low absolute distortion caused by scrambling. If there are however two or more pilot sequences the relative error strongly increases as besides 'scattering out' also 'scattering in' has to be regarded: The intrusion of energy from other codes also happens at time (slot) positions where the respective 'target' pilot signal is weak or even vanishes, causing an error signal, which can be much higher than the 'pure' signal.

If MIMO is configured for more than one data stream these streams will interfere with each other with strength depending on the correlation of the different physical channels. But this interference can effectively be reduced by interference cancellation as all data streams have the same origin and therefore are subject of the same channel conditions. Furthermore, the receiver has to decode all data streams anyway so that the additional effort of cancellation is limited. As interference cancellation is re-applying the channel transformation to the already detected data, any error in the estimated channel information takes effect three times, namely at first equalization, at re-application of the channel operation in the course of the interference cancellation and at re-equalization of the equivalent single stream data. Any error in channel estimation therefore acts non-linearly on the overall performance of the system.

The scrambling distortion is mainly significant for very

high data rates where the respective user equipment (UE) is served by the base station with all available resources. This means that intra-cell interference (originating from other users) is small, mainly caused by pilot and control channels, which are known to the UE and can be suppressed. Furthermore, these mentioned high data rates are achievable under quite good radio conditions only, i.e., also the inter-cell interference must be small. Therefore, a correlation-based estimator might be sufficient in that range of interest but the impact of systematic scrambling induced distortion should be removed.

As the receiver knows the scrambled pilot symbol patterns as they are fed to the transmitting antennas it is possible to calculate the distortion induced by scrambling analytically as a function of the channel impulse response. Inverting this function allows to extend the correlation algorithm such as to take into account the distortion, leading to an estimated channel impulse response being exact with respect to scrambling effects. Although a matrix operation is required to solve the resulting equation system the corresponding matrix elements depend only on the scrambling vector and can hence be calculated once at the initial phase. As long as the scrambling code is not changed, i.e., the UE does not change the cell, these elements remain stable.

The remainder of this paper is organized as follows: In Section II we give a brief description of MIMO inter-stream interference and its cancellation as this is the main application of the proposed improvement in channel estimation. In Section III we then introduce the signal model with the basic equations. They are required to describe the data transfer through the physical channels to be measured by the UE and will be used in Section IV for formulating the analytical solution for scrambling-exact channel estimation. Section V then demonstrates the gain of the proposed algorithm with some results from link level simulation, followed by a discussion of an efficient implementation of the algorithm in Section VI. After explaining why the deviation of scrambling from orthogonality has no impact on data detection itself in Section VII, we finally draw our conclusion in Section VIII.

II. Cancellation of MIMO Inter-Stream Interference

In W-CDMA HSDPA, the maximum possible data rate has been and will be increased over the years especially by using modulation schemes of higher order such as 16QAM or 64QAM and new modes of operation [10], [11]. The latter can be divided into two classes. The first one is extending the bandwidth by using two or even more carriers. In that case, the procedures on physical layer remain more or less untouched and are hence beyond our focus here. The second one is the introduction of MIMO with transferring several data streams in parallel using both the same physical resources as well as the same spreading codes

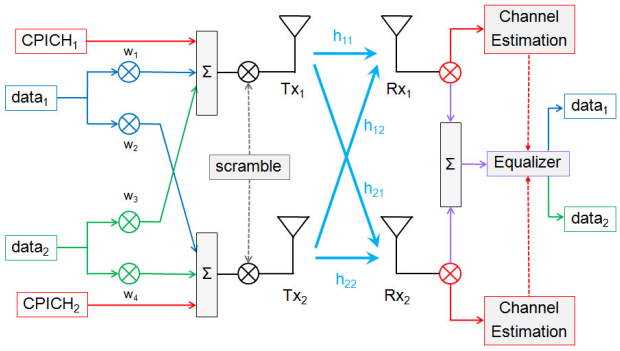


Figure 1. Schematic overview about dual stream transmission in W-CDMA HSDPA MIMO.

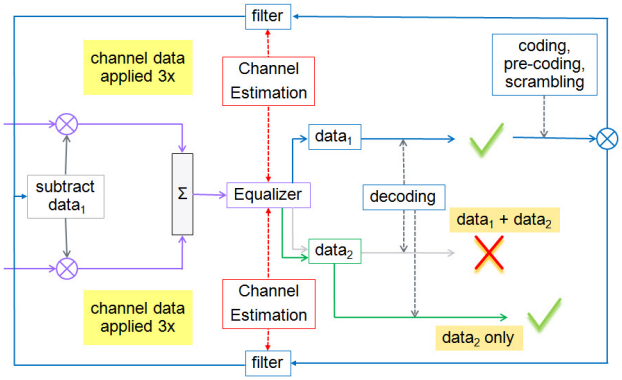


Figure 2. Schematic overview about cancellation of inter-stream interference.

[12]. Fig. 1 gives a schematic overview about MIMO dual stream transmission.

Good channel estimation plays a significant role for effective data transmission, especially if high data rates shall be reached. This is achieved by transmitting pilot patterns with defined power, which are used then by the receiver to estimate the single channel coefficients. Dual stream transmission requires at least two antennas on both transmitter and receiver side. To enable the receiver to measure each physical channel separately, two different pilot signals are sent. In HSDPA each one is fed to one of the transmitting antennas [13], whereas each data stream is distributed to both transmitting antennas with different phase factors, the pre-coding weights. A pre-coding codebook with four entries is defined and the UE signals the index of the pre-coding values, which yield the highest received power. The pre-coding vector for the second data stream is then taken orthogonal to the one of the first stream in order to allow the separation of the streams by the receiver. The received power of the second stream usually is significantly lower due to more destructive interference of the beams from the two transmitting antennas. For that reason, first and second stream often are refered as strong and weak one. Further details are not important for the discussion here but

can be found in [12].

In conventional LMMSE equalizers as described, e.g., in [14] both pre-coding weights and channel coefficients are needed to determine proper weights for calculation of soft values of the received symbols, which then are passed to the decoder. Although these equalizer weights are non-linear in the channel impulse response the weight calculation is the only place in the decoding chain where the estimated channel coefficients are required.

The separation of the streams on receiver side is possible due to diversity of the physical channels between each transmitting and receiving antenna either spatially or due to different polarization properties of the affected antennas. If the physical channels are correlated this separation of the streams is feasible only to a certain extend and a substantial inter-stream interference remains after equalizations, which mainly effects the weaker of the two data streams. To keep this stream decodable anyway requires the application of a modulation and coding scheme with rather low transport block size.

If, however, the first data stream is decoded correctly, the complete information for re-constructing its shape as it reaches the receiver antennas is available at the receiver, including the coefficients of the physical channels. It is therefore possible to subtract this contribution from the overall incoming signal, leaving the stripped signal containing only the signal of the second stream, and of course noise and interference beyond inter-stream interference. This cancellation of inter-stream interference thus allows higher data rates also for the weaker data stream.

From the schematical display in Fig. 2 it becomes obvious that now the estimated channel coefficients are not applied only once but three times, namely for the first calculation of the equalizer weights, for the re-construction of the stronger signal as it reaches the antennas, and for the re-calculation of the weights used for equalizing the remaining weaker stream, which now requires single stream weights. As long as the two data streams are not orthogonal at the reciever side, which is usually the case if interference cancellation is executed, application of dual stream weights implicitly assigns within the equalizer equations a part of the stream power to the other data stream, which is, however, no more present. This both reduces the available signal over noise ratio and distorts the signal, i.e., the data detection suffers significantly [15].

The increased impact of channel estimation errors in presence of interference cancellation becomes obvious when looking to the equation for the received soft values $y_{2,\mathrm{IC}}$ of the weaker second data stream at the equalizer output (see, e.g., [16]).

$$y_{2,\mathrm{IC}} = w_{2,\mathrm{single}} \left[r - (1-\delta) \cdot h \cdot s_1 \right]. \qquad (1)$$

To the re-spread and re-scrambled signal of the first data stream, s_1, the channel impulse response h is re-applied. Errors in channel estimation are accounted for by the correction δ. As the main part of the overall received signal r originates from the stronger first data stream, both terms in the bracket are of the same order of magnitude and $y_{2,\mathrm{IC}}$ is therefore quite sensitive to errors in the channel estimation with leading order of δ. The additional contribution of δ implicitly contained in $w_{2,\mathrm{single}}$ is present without interference cancellation, too.

Of course it would be much easier to use directly the stronger signal at the equalizer output instead of re-coding the decoded one, especially as this would reduce the detection delay. The benefit of such a procedure however is quite limited because such a signal still contains all noise and interference contributions and not only errors from channel estimation. It can even happen that the manipulated signal obtained this way is worse than without interference cancellation at all.

In Section V we will show that the performance of the weaker data stream experiences a huge gain from improved channel estimation, but also the stronger data stream as well as the weaker one in absence of interference cancellation benefit from this procedure.

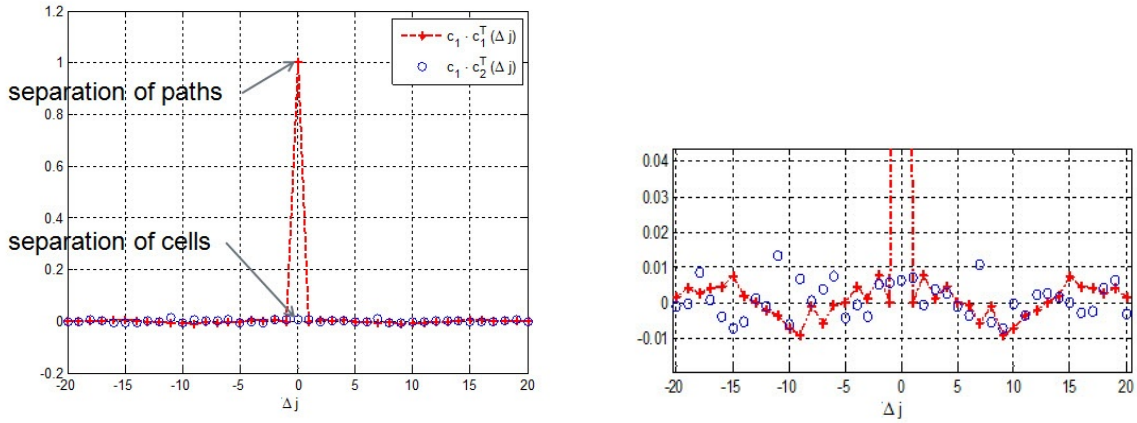
Table I
NAMING CONVENTIONS FOR IMPORTANT VARIABLES

Received signal (single antenna)	$r(l)$
Sent signal (index i for Tx antenna)	$s_i(l)$
Channel impulse response (index i for Tx antenna)	$h_i(l)$
Scrambling code element of sequence j at integer chip index n	$c_j(n)$
Channel length	$M+1$
Spreading sequence length	N_{SF}
Oversampling factor	N_{OSF}
Number of symbols required for CPICH orthogonality	N_{ortho}
Number of symbols required for CPICH orthogonality, measured in chips	$N = \frac{\text{chips per slot}}{N_{\mathrm{ortho}}}$

III. BASIC CHANNEL PILOT EQUATIONS

Throughout the rest of the paper, we use the naming conventions listed in Table I. We assume that two transmitting antennas are present, each fed with a separate pilot sequence, namely s_1 and s_2. On chip level, i.e., after spreading, the signals are then chip-wise multiplied with the elements of a scrambling vector $\mathbf{c}$, which has the length of a frame, i.e., 38400 chips [17]. The scrambling sequence is known also to the receiving UE, which can hence de-scramble the received signal r by multiplication with the hermitian scrambling vector, $\mathbf{c}^{\dagger}$.

The W-CDMA scrambling sequences are constructed from the pseudorandom complex-valued Gold sequences [17] and hence their autocorrelation is quasi-orthogonal (see,



(a) (b)

Figure 3. Scalar product of scrambling vector with itself (red) and with another (blue) scrambling vector as function of shifted index.

e.g., [18]. This operation preserves both amplitude and phase to a high extend if the applied scrambling vector matches the one used on the transmitter side without shift in the chip index as can be seen from the red curve at $j = 0$ in Fig. 3a. On the other hand, de-scrambling with another scrambling sequence (blue curve in Fig. 3a) or the same one but shifted in the chip index yields an correlation result close to zero. If really orthogonal instead of quasi-orthogonal codes would be used non-matching de-scrambling would provide exactly vanishing results. Scrambling is therefore a powerful means for both separation of signals from different cells, which use different scrambling codes, as well as for separation of contributions of the same base station but propagating via differently delayed paths, i.e., application of the same but time-shifted scrambling sequence.

Fig. 3b shows the error caused by the deviation of scrambling vectors from orthogonality in more detail. Although this error is in the range of a few percents only it can cause a degradation of system performance, especially if advanced receivers are used.

As we want to include oversampling and raised cosine filtering in our description, s and r are defined in sub-chip time steps l. For a single receiving antenna, the received signal is then given by (see, e.g., [18], Section 1.3):

$$r(l) = \sum_{k=0}^{N_{\text{OSF}} \cdot M} c\left(\left\lfloor \frac{l-k}{N_{\text{OSF}}} \right\rfloor\right) \left[h_1(k) s_1\left(\frac{l-k}{N_{\text{OSF}}}\right) + h_2(k) s_2\left(\frac{l-k}{N_{\text{OSF}}}\right)\right], \quad (2)$$

where $\lfloor \cdots \rfloor$ denotes the interger part of the argument.

Usually, the raised cosine filtering is counted for in the channel impulse response h. The signal $s_i, i \in \{1, 2\}$ upsampled with oversampling factor N_{OSF} is then non-zero only for integer arguments and hence only M terms contribute to the summation in (2). For further receiving antennas the same equation holds with the corresponding channel impulse responses.

Although the conventional correlation approach of channel estimation, which de-scrambles the received signal ignoring the error made by this quasi-orthogonal operation (see, e.g., [2]), allows to separate both signals from different cells with different scrambling sequences c_1 and c_2 as well as to separate different delays Δj of the signals from the same cell, it contains a systematic error. This error is visible in Fig. 3 where the scalar product of a scrambling vector with itself (red) and with another scrambling vector (blue) is displayed. In both cases, the second vector is circularly shifted in its index Δj. In case of real orthogonality, only the product of a scrambling vector with itself and matching indices (red cross at $\Delta j = 0$) would yield a non-vanishing result.

In order to find an exact solution for de-scrambling, we build the projection of the received signal to the scrambling vector but keep all terms resulting from deviation of scrambling from orthogonality:

$$\begin{aligned} &\sum_{l=l_0+1}^{l_0+N_{\text{SF}}} c^{\dagger}(l)\, r(l+d) = \\ &= \sum_{l=l_0+1}^{l_0+N_{\text{SF}}} \sum_{k=0}^{N_{\text{OSF}} \cdot M} c^{\dagger}(l) c\left(\left\lfloor \frac{l-k+d}{N_{\text{OSF}}} \right\rfloor\right) \cdot \\ &\quad \left[h_1(k) s_1\left(\frac{l-k+d}{N_{\text{OSF}}}\right) + h_2(k) s_2\left(\frac{l-k+d}{N_{\text{OSF}}}\right)\right] \approx \\ &\approx h_1(d) s_1(l_0) + h_2(d) s_2(l_0). \end{aligned} \quad (3)$$

The approximation in the last step applies the simplification of the correlation approach.

In (3), we introduced a start index l_0 for summation over chips in order to cope with handling of more than one symbol. For the first symbol, $l_0 = 0$, for the second

one $l_0 = N_{SF}$, the spreading sequence length, and higher symbol indices are related to the corresponding integer multiples of N_{SF}.

Although the scrambling vector $\mathbf{c}$ and its index l operate on chips, both the received data $\mathbf{r}$ and the channel impulse response function $\mathbf{h}$ are defined on sub-chip level and thus d and k take into account oversampling. During the upsampling operation, the data vectors $\mathbf{s}_i, i \in \{1, 2\}$, are extended according to the oversampling factor by filling the new elements with zeros. This means that all terms on the right hand side of (3) with $(l - k + d)/N_{\text{OSF}} \notin \mathbb{Z}$ vanish because the corresponding $s_i = 0$. If we now define

$$\begin{aligned}
\kappa_{l_0,j}(d,k) &\equiv \kappa_{l_0,j}(d-k) = \\
&= \sum_{l=l_0+1}^{N_{\text{OSF}}\cdot N_{\text{SF}}} c^\dagger(l) c\left(\left\lfloor \frac{l-k+d}{N_{\text{OSF}}} \right\rfloor\right) s_j\left(\frac{l-k+d}{N_{\text{OSF}}}\right) \\
\kappa_{l_0,j}(d-k) &= 0 \quad \forall d-k \notin \mathbb{Z} \\
\rho_{l_0}(d) &= \sum_{l=l_0+1}^{N_{\text{SF}}} c^\dagger(l) r(l+d) \qquad (4)
\end{aligned}$$

for a channel of length $M+1$ measured in chip duration equation (3) can be written as

$$\rho_{l_0}(d) = \sum_{k=0}^{N_{\text{OSF}}\cdot M} \left[h_1(k)\kappa_{l_0,1}(d-k) + h_2(k)\kappa_{l_0,2}(d-k)\right]. \qquad (5)$$

The correlative de-scrambling is contained in this formulation as special case, namely setting $\kappa(d-k) = \delta_{d,k}$. The channel coefficients are then directly given by ρ.

We are now able to split the summation into two parts, namely one counting the chips and the other stepping through the oversampling within each chip. With the definitions

$$\begin{aligned}
d &= N_{\text{OSF}} \left\lfloor \frac{d}{N_{\text{OSF}}} \right\rfloor + d \mod N_{\text{OSF}} = \hat{d} + \check{d} \\
\hat{d} &= \left\lfloor \frac{d}{N_{\text{OSF}}} \right\rfloor \qquad (6) \\
\check{d} &= d \mod N_{\text{OSF}}
\end{aligned}$$

and assuming that l is an integer multiple of N_{OSF} (generally spoken, oversampling can be accounted for in d) we can write finally

$$\begin{aligned}
&\rho_{l_0}(\hat{d}+\check{d}) = \\
&= \sum_{k=0}^{M} \sum_{j=0}^{N_{\text{OSF}}-1} \Big[h_1(N_{\text{OSF}}k+j)\kappa_{l_0,1}(\hat{d}+\check{d}-N_{\text{OSF}}k-j) + \\
&\qquad h_2(N_{\text{OSF}}k+j)\kappa_{l_0,2}(\hat{d}+\check{d}-N_{\text{OSF}}k-j)\Big] = \\
&= \sum_{k=0}^{M} \Big[h_1(N_{\text{OSF}}k+\check{d})\kappa_{l_0,1}(\hat{d}-N_{\text{OSF}}k) + \\
&\qquad h_2(N_{\text{OSF}}k+\check{d})\kappa_{l_0,2}(\hat{d}-N_{\text{OSF}}k)\Big]. \qquad (7)
\end{aligned}$$

In (7), we have separated integer and oversampled parts in the sum of the right hand side of (5) by splitting

$$k = N_{\text{OSF}} \left\lfloor \frac{k}{N_{\text{OSF}}} \right\rfloor + j \overset{\left\lfloor \frac{k}{N_{\text{OSF}}} \right\rfloor \to k}{\longrightarrow} N_{\text{OSF}}k + j, \qquad (8)$$

i.e., the summation over k is restricted to run from 0 to M whereas the oversampled indices are collected in j, which runs from 0 to N_{OSF} and finally is executed according to the definition given in (4).

IV. Scrambling-Exact Channel Estimation

In [8], there are defined different configurations for the common pilot channel (CPICH) in W-CDMA systems. For the usage in combination with MIMO, two of them are relevant, namely either using only the primary CPICH spread with code c_{SF1} of length 256 and orthogonal bit sequences on the two transmitting antennas, or using the primary CPICH on the first and a secondary CPICH with code c_{SF2} on the second transmitting antenna. The secondary CPICH than must use the same bit sequence as the primary CPICH. The same scrambling code for both cases is used in any case.

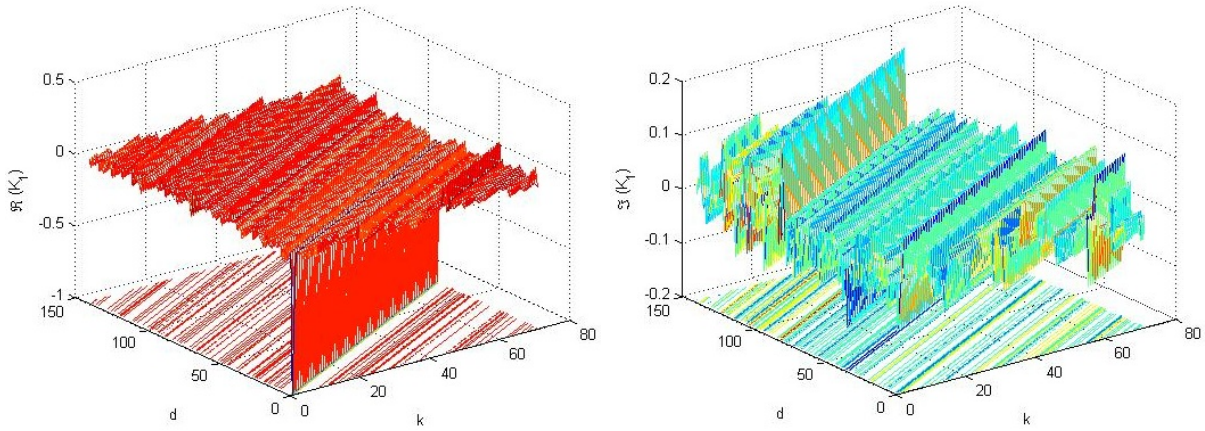
If two adjacent symbols are merged both options can formally be described by a virtual spreading with length 512 where the effective spreading codes can be defined in the first case as $[c_{\text{SF1}}\, c_{\text{SF1}}]$ and $[c_{\text{SF2}}\, \text{inv}(c_{\text{SF1}})]$, and in the second case as $[c_{\text{SF1}}\, c_{\text{SF1}}]$ and $[c_{\text{SF2}}\, c_{\text{SF2}}]$. As long as the channel estimation algorithm finally uses any filtering procedure over an even number of pilot symbols, both options are equivalent with respect to the influence of scrambling. We will restrict therefore the following investigations to the configuration with primary CPICH only and orthogonal bit sequences on the two transmitting antennas.

A. Single Symbol Channel Estimation

If we want to calculate $M+1$ coefficients for both channels, we need $2 \cdot (M+1)$ equations, i.e.,

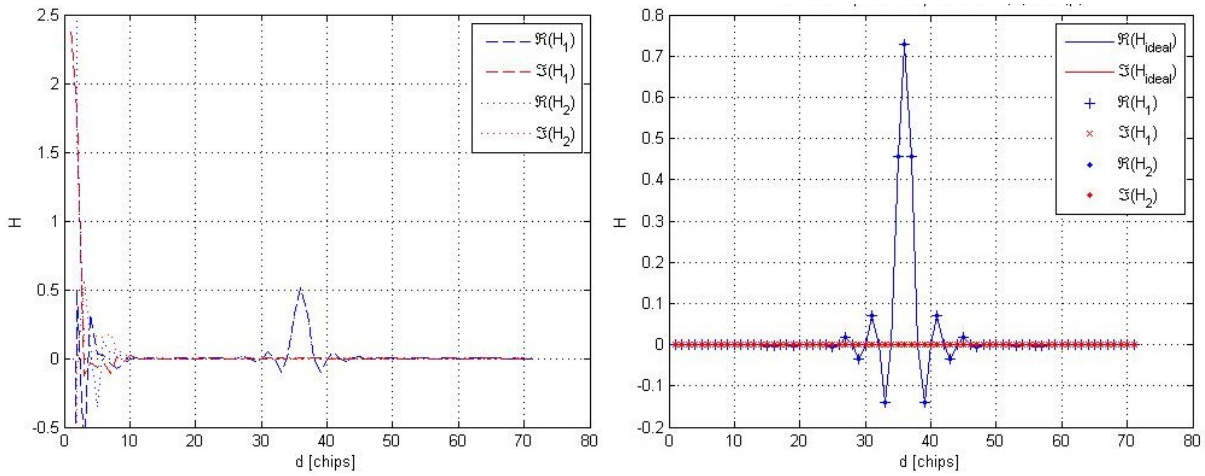
$$\begin{pmatrix} \rho_{l_0}(N_{\text{OSF}}\cdot 0+\check{d}) \\ \vdots \\ \rho_{l_0}(N_{\text{OSF}}\cdot M+\check{d}) \\ \vdots \\ \rho_{l_0}(N_{\text{OSF}}\cdot(2M+1)+\check{d}) \end{pmatrix} = \mathbf{K}_{l_0} \begin{pmatrix} h_1(N_{\text{OSF}}\cdot 0+\check{d}) \\ h_2(N_{\text{OSF}}\cdot 0+\check{d}) \\ \vdots \\ h_1(N_{\text{OSF}}\cdot M+\check{d}) \\ h_2(N_{\text{OSF}}\cdot M+\check{d}) \end{pmatrix}, \qquad (9)$$

where the pilot scrambling matrix $\mathbf{K}_{l_0}$ is defined as



(a) (b)

Figure 4. Real (a) and imagionary (b) part of odd columns of pilot scrambling matrix K for a single tap channel including raised cosine filtering.



(a) (b)

Figure 5. Channel Impulse Response H averaged over all symbols of a frame (a) and calculated with averaged K (b).

$$\mathbf{K}_{l_0} = \tag{10}$$

$$\begin{pmatrix} \kappa_{l_0,1}(0) & \kappa_{l_0,2}(0) & \dots & \kappa_{l_0,1}(-M) & \kappa_{l_0,2}(-M) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \kappa_{l_0,1}(M) & \kappa_{l_0,2}(M) & \dots & \kappa_{l_0,1}(0) & \kappa_{l_0,2}(0) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \kappa_{l_0,1}(2M+1) & \kappa_{l_0,2}(2M+1) & \dots & \kappa_{l_0,1}(M+1) & \kappa_{l_0,2}(M+1) \end{pmatrix}.$$

If the scrambling sequences would be completely orthogonal, the pilot scrambling matrix (10) would contain non-zero elements ($\kappa_j(0)$) in the first $M+1$ rows only, i.e., rows $M+1 \dots 2M+1$, would not contribute at all to the solution thus making $\mathbf{K}$ un-invertible. The accuracy of these equations is therefore small even if the non-orthogonality of scrambling is taken into account. To overcome this limitation one has therefore either to average over at least as many symbols as there are pilot sequences (see Section IV-B), or the approach has to be extended such as to obtain equations for each pilot symbol (see Section IV-C). This becomes obvious when looking to the pilot scrambling matrix for a single-tap channel including root raised cosine filtering as displayed in Fig. 4.

The upper half of $\mathbf{K}$ shows large real entries on the main diagonal related to κ with argument 0, whereas the other elements are dominated by the small scrambling correlation coefficients with arguments different from 0. As expected, the calculation of the channel impulse response fails at least partly even if afterwards an averaging over all symbols of a frame is performed. This is shown in Fig. 5a for an AWGN channel with raised cosine filter.

The channel coefficients (at sampling positions) are finally

given by

$$\begin{pmatrix} h_1(0) \\ h_2(0) \\ \vdots \\ h_1(M) \\ h_2(M) \end{pmatrix} = \mathbf{K}^{-1}_{\text{single},l_0} \begin{pmatrix} \rho_{l_0}(0) \\ \vdots \\ \rho_{l_0}(M) \\ \vdots \\ \rho_{l_0}(2M+1) \end{pmatrix}. \quad (11)$$

B. Averaging Pilot Scrambling Matrix

The inaccuracy of the channel impulse response calculated for a single symbol is caused by $\mathbf{K}$ being badly conditioned. This can be improved by averaging both $\mathbf{K} \rightarrow \langle \mathbf{K} \rangle$ and $\rho \rightarrow \langle \rho \rangle$. Although the general shape of the pilot scrambling matrix is preserved by this operation the conditioning is now about 10 orders of magnitude better and $\langle \mathbf{K} \rangle_{\text{frame}}$ becomes invertible. For an AWGN channel with raised cosine filter it provides a perfectly estimated channel impulse response, see Fig. 5b.

C. Twin Symbol Channel Estimation

The averaging procedure discussed in Section IV-B assumes that $\mathbf{K}$ and ρ are statistically independent and the averaging therefore factorizes and can be executed independently. Inversion is then executed with the averaged pilot scrambling matrix. The shown result implies that this assumption is correct at least for a single tap channel including root raised cosine filter.

There is, however, an alternative approach, which does not need the above restriction: Both pilot scrambling matrix and de-spread received signal vector are constructed using two symbols with different pilot sequences,

$$\begin{pmatrix} \rho_{l_0}(0) \\ \vdots \\ \rho_{l_0}(M) \\ \rho_{l_1}(0) \\ \vdots \\ \rho_{l_1}(M) \end{pmatrix} = \mathbf{K}_{\text{twin},l_{0,1}} \begin{pmatrix} h_1(0) \\ h_2(0) \\ \vdots \\ \vdots \\ h_1(M) \\ h_2(M) \end{pmatrix}, \quad (12)$$

where the pilot scrambling matrix $\mathbf{K}_{\text{twin}}$ now is defined as

$$\mathbf{K}_{\text{twin},l_{0,1}} = \quad (13)$$
$$\begin{pmatrix} \kappa_{l_0,1}(0) & \kappa_{l_0,2}(0) & \dots & \kappa_{l_0,1}(-M) & \kappa_{l_0,2}(-M) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \kappa_{l_0,1}(M) & \kappa_{l_0,2}(M) & \dots & \kappa_{l_0,1}(0) & \kappa_{l_0,2}(0) \\ \kappa_{l_1,1}(0) & \kappa_{l_1,2}(0) & \dots & \kappa_{l_1,1}(-M) & \kappa_{l_1,2}(-M) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \kappa_{l_1,1}(M) & \kappa_{l_1,2}(M) & \dots & \kappa_{l_1,1}(0) & \kappa_{l_1,2}(0) \end{pmatrix}.$$

In the following, these symbols will be called twin symbols if the two symbols are adjacent. The channel coefficients are then given by

$$\begin{pmatrix} h_1(0) \\ h_2(0) \\ \vdots \\ \vdots \\ h_1(M) \\ h_2(M) \end{pmatrix} = \mathbf{K}^{-1}_{\text{twin},l_{0,1}} \begin{pmatrix} \rho_{l_0}(0) \\ \vdots \\ \rho_{l_0}(M) \\ \rho_{l_1}(0) \\ \vdots \\ \rho_{l_1}(M) \end{pmatrix}. \quad (14)$$

Now, all rows and columns of the pilot scrambling matrix contain (at least) one main element as displayed in Fig. 6. The real part shows two bars, one completely in the lower half volume, the other in both half volumes. Whereas the former belongs to equal pilot symbols for both channels, the latter results from symbols of the second pilot being inverse to the ones of the first pilot.

The AWGN channel impulse response including raised cosine filter calculated from the twin pilot correlation matrix again matches perfectly with the ideal channel impulse response, not only when averaged over all symbols of a frame (see Fig. 7), but also for each twin symbol. This means that no averaging is required in the absence of noise.

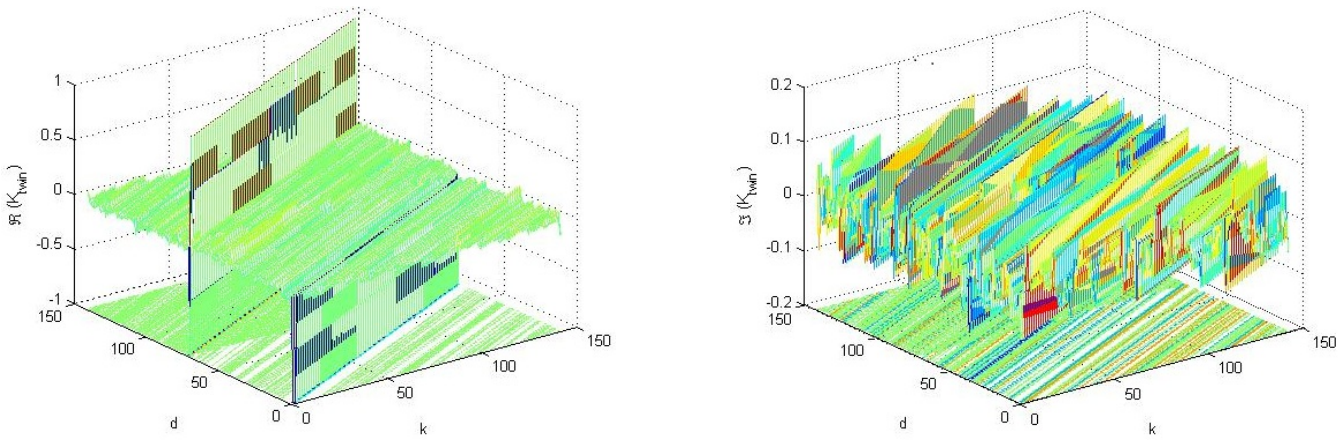
V. Link Level Simulations

In order to demonstrate the benefit of the scrambling-exact channel estimation we executed link level simulations for a single UE moving with 3km/h and assuming a channel of type Pedestrian A as it is defined, e.g., in [19], table 1.2.2.2, Channel A. Two data streams were transmitted in MIMO mode and for both the modulation and coding scheme corresponding to entry 10 in the CQI mapping table K [12] was used. The transmitting antennas were assumed to be mounted crosswise diagonal (X), the receiving antennas crosswise perpendicular vertical and horizontal (+) and crosstalk between the polarization directions was allowed. The basic model is displayed in Fig. 8.

The general relation between signals $\mathbf{s}_i, i \in \{1, 2\}$ sent from antenna i and the signal $\mathbf{r}_k$ received at antenna k is then given by

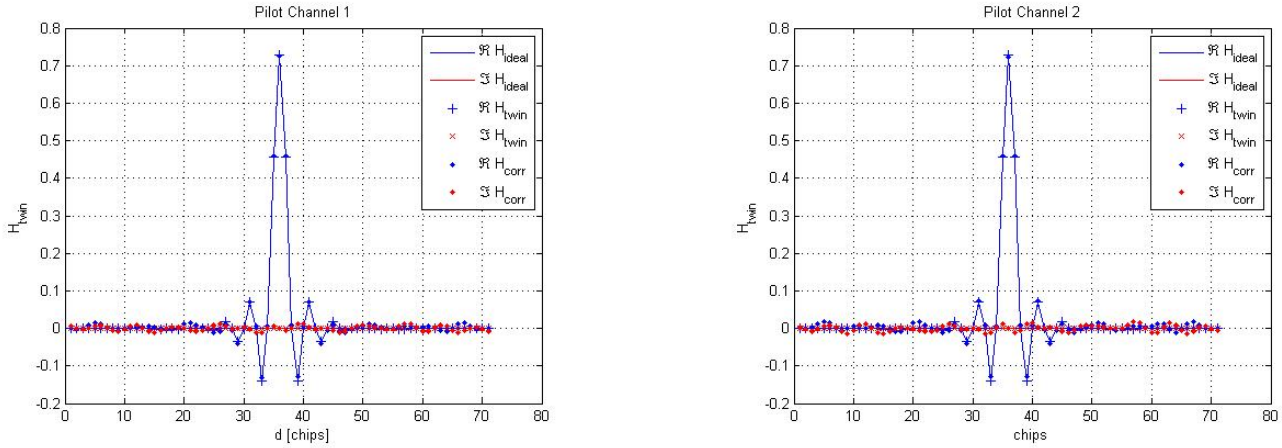
$$\mathbf{r}_k = \begin{pmatrix} \cos\theta_k & \sin\theta_k \end{pmatrix} \begin{pmatrix} h_{Hk}\sqrt{1-\kappa^2_{HV}} & h_{Vk}\kappa_{VH} \\ h_{Hk}\kappa_{HV} & h_{Vk}\sqrt{1-\kappa^2_{VH}} \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{pmatrix}, \quad (15)$$

with $\phi = \pi/4$ for X-mounted transmitting antennas and $\theta = \{\pi/2, 0\}$ for receiving antennas with + configuration. The parameters κ_{HV} and κ_{VH} take into account crosstalk from horizontal to vertical and from vertical to horizontal polarization direction [20], respectively. Besides crosstalk the channels were assumed to be uncorrelated.



(a) (b)

Figure 6. Real (a) and imaginary (b) part of Pilot scrambling matrix K_{twin} for one twin symbol.



(a) (b)

Figure 7. Twin Channel Impulse Response averaged over all symbols, compared with conventional correlation result and ideal case.

The block error rate (BLER) for each data stream was determined as a function of SNR. As measure for performance serve the SNRs for a working point (WP) at 10% BLER. The simulator operated on sub-chip level with oversampling factor $N_{\text{OSF}} = 2$, and also raised cosine filtering was included.

To demonstrate the enhanced impact of channel estimation errors in case of cancelling inter-stream interference, Fig. 9 shows simulation results with channel estimation via the correlation approach (denoted with *corr*) and with scrambling-exact (*twin*) channel estimation for both interference cancellation (*IC*) switched on and off. The respective SNR of the working points measured in dB is given listed in the legend.

Regarding the results for the first, stronger data stream displayed with solid lines it can be seen that interference cancellation has practically no impact on the performance, the SNR difference of the working point is 0.1dB, whereas scrambling exact channel estimation gains about 1.5dB. Without interference cancellation, the situation is similar also for the second, weaker data stream: The benefit of scrambling-exact channel estimation for the working point is about 1.5dB, but a really huge gain is achieved if interference cancellation and scrambling-exact channel estimation are combined: In this case, the working point is shifted overall by more than 5dB. This result confirms the expectation of Section II where we explained the increased impact of errors in channel estimation when the contribution of the correctly decoded data stream is removed from the incoming signal.

VI. Runtime-Efficient Implementation

From the explanation in Section III it is clear that a scrambling-exact channel estimation has to be paid with increased implementation and runtime effort. Instead of

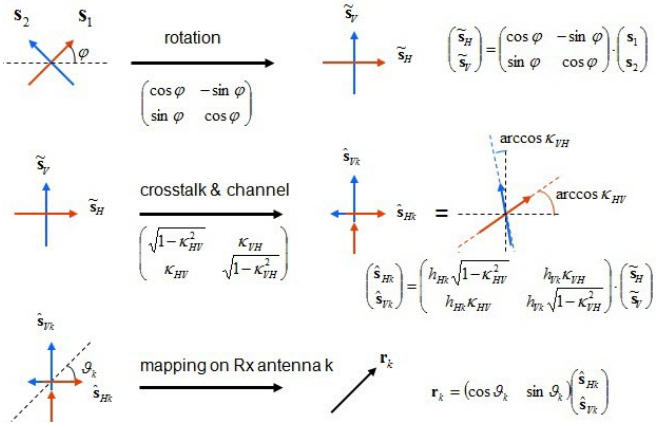


Figure 8. Channel model including crosstalk between horizontal and vertical polarized beams.

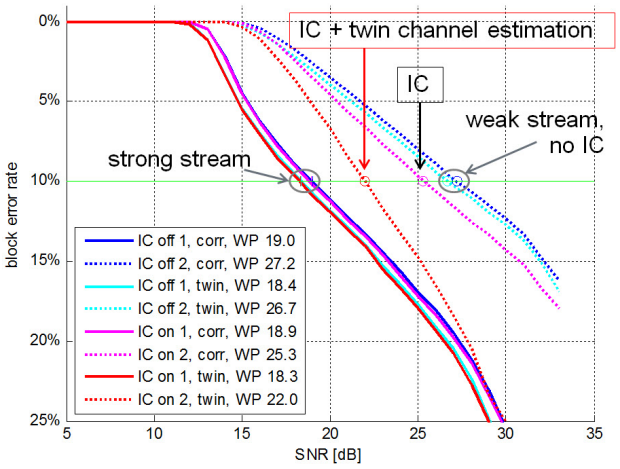


Figure 9. Link level simulation results with simple correlation and scrambling-exact channel estimation with and without interference cancellation.

solving (5) in the correlative limit,

$$\rho_{l_0}(d) = h_1(d) + h_2(d), \tag{16}$$

the elements κ of the pilot scrambling matrix $\mathbf{K}$ defined in (4) have to be calculated and the received signal convoluted with the scrambling vector to be divided by $\mathbf{K}$.

Fortunately, the pilot symbols s are fix and the scrambling code elements c are constant at least within a cell. This allows executing the more expensive part, the calculation of the matrix elements, in an initialisation phase during the handover. Please note that the pilot scrambling matrix is required only for decoding the HSDPA payload, not for the control channels as the latter are not transmitted in MIMO but in legacy SISO mode even if MIMO is active for the respective UE.

The cheaper part is the division by $\mathbf{K}$, a matrix of size $2(M+1)\times 2(M+1)$, which has to be executed permanently,

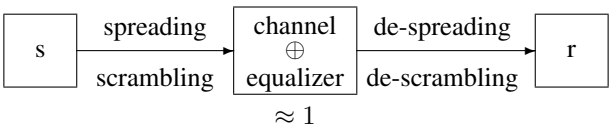


Figure 10. Spreading and de-spreading chain for transmitted data.

i.e., during each update of the channel estimation as it is applied to the varying received pilot signal. As the dimensions of $\mathbf{K}$ are comparable to the size of the LMMSE equalizer matrix, which also contains matrix division and several matrix multiplications, the equalization procedure can be regarded as upper boundary for estimating the runtime effort of this operation.

In principle it is possible to calculate the inverse of $\mathbf{K}$ in advance, too, but this brings only minor benefit as matrix division and matrix multiplication are of similar complexity for good conditioned matrix inversion. Even more important is, however, the fact that the scrambling sequence c has a length of 38400 [17] (number of chips in one frame) and the inverse of all potential excerpts of it have to be prepared if they shall be pre-calculated. Even if one restricts to solving the channel estimation equation always within the twin symbol boundaries, which have a length of 512 chips, 75 different matrices would have to be stored.

In the calculation of the matrix elements κ of (4) two types of arguments occur, namely the summation index l and the shifted index $(l+d-k)/N_{\mathrm{OSF}}$. Including the raised cosine filter in the channel impulse response leaves the upsampled but not yet filtered signal for s, which is non-zero only for integer arguments, reducing thus the number of elements contributing to the sum to $M+1$. The elements of κ hence are non-zero only on chip-level,

$$\begin{aligned} \kappa_{l_0,j}(k) &= \sum_{l=l_0+1}^{N_{\mathrm{SF}}} c^{\dagger}(l)c(l+k)s_j(l+k) \\ &= \sum_{l=l_0+1}^{N_{\mathrm{SF}}} c^{\dagger}(l)\sigma_j(l+k), \\ \sigma_j(i) &\equiv c(i)s_j(i). \end{aligned} \tag{17}$$

Now, the elements for the required sequence of the channel impulse response including the raised cosine filter, $k \in -M \ldots M$, can be calculated within one loop by just circular shifting the index of σ by one step to the right with increasing k. The matrix $\mathbf{K}$ is then constructed by cutting out line-wise the required part of κ.

VII. Scrambling Impact on Data Detection

The elaboration in the previous sections has shown the significant impact of scrambling on channel estimation and

hence rises the question how quasi-orthogonality of scrambling affects data detection. To discuss this in more detail we will start with the data at the equalizer output and assume that the data streams are well separated and the transmitted signal before upsampling is re-constructed. We can then restrict the discussion to the process chain as displayed in Fig. 10.

This operation chain can formally be written as

$$r(l) = \sum_{k=0}^{N_{\mathrm{SF}}-1} c^{\dagger}(l+k)c(l+k)s(l+k). \qquad (18)$$

From (18) we see that de-scrambling the data only makes use of the normality but not of the quasi-orthogonal property of scrambling as only coefficients c at the same time index $l+k$ occur. De-scrambling the data is therefore not affected by the deviation of scrambling sequences from orthogonality. Of course quasi-orthogonality is required also for data detection in order to remove interference from other cells but this cannot be included correctly here as otherwise both scrambling sequences and signal strengths of all potential interferers would have to be known. In contrast to the two pilot signals, interfering signals propagate along different physical channels, which are especially different from the channel relevant for both data and pilot signals of the UE of interest. Therefore, it can be expected that superposition of different interferers and equalisation with channel coefficients not matching these physical channels the interfering signals propagate through will lead to a further whitening of these contributions so that no systematic error persists the interfering signals from other cells.

VIII. Conclusion

The usual procedure for estimating the radio channels discards the small non-orthogonality of the autocorrelation function of the pseudorandom scrambling sequence. As long as only one pilot signal is transmitted this effect is negligible as the error is more or less proportional to the strength of the received pilot signal at the respective chip position. Delayed multi-path taps bearing no pilot signal for the respective chip are not affected at all.

The usage of multiple transmitting antennas, e.g., in STTD or MIMO systems, allows to benefit from spatial and polarization diversity, which however requires to measure all physical channels from each transmitting to each receiving antenna. This is possible only if different pilot patterns (or different linear combinations of them generated by different pre-coding) are sent from the different transmitting antennas. Ignoring deviation from orthogonality then causes that energy scattered by scrambling out of a certain pilot pattern will appear in another pilot pattern. This contribution is no more proportional to the signal strength of the respective pilot signal at the regarded chip. Also taps in fact containing no pilot signal from the investigated chip position will experience a signal contribution causing a huge relative error.

It has to be emphasized that this error is present even in case of pilot channels being fully separated, e.g., by cross polarized wave fronts with each front bearing exactly one pilot and receiving antennas oriented along the polarization directions, as long as this fact is not known a priori on receiving side: The estimator has to assume that signals of both pilots are present at each receiving antenna and hence misinterprets fractions scattered from the 'real' pilot by scrambling as contributions of the other pilot. Merely a single pilot being present combined with the knowledge about this fact on receiving side avoids this complication.

Due to the importance of channel estimation for reaching high data rates, there are published many proposals for improving the basic correlation approach, e.g., by filtering the primary channel estimates [4], cancelling inter-antenna interference [5] or extending the RAKE correlation approach to a LMMSE one [3] and thus taking into account all kinds of interference. All of these suggestions can basically be extended by scrambling-exact dealing the pilot patterns as proposed in [1] even if we have investigated here only the simple RAKE correlation approach, including appropriate filtering as described in [4], in the simulator.

Although the scrambling induced error does not occur in data detection itself and is negligible in channel estimation in many cases, high data rates in combination with MIMO and transmission of two data streams require more accurate channel estimation. This can be reached by taking into account the scrambling sequence exactly in the channel estimation algorithm. As in the correlation case, two symbols (in case of two pilot signals) with different pilot code elements are required to resolve the channels from both transmitting antennas. Therefore, a twin symbol pilot scrambling matrix is introduced with its inverse applied to the received signal providing the channel impulse response.

The numerical effort of this approach is for sure higher than the conventional correlation method. This effort however can at least partly be shifted to an initialization phase because the pilot scrambling matrix only depends on the scrambling sequence, the pilot patterns and the maximum length of the channel but not on any quantities varying with time. The complexity of an efficient implementation is then comparable to the one of an LMMSE equalizer. Mainly the second (weaker) data stream benefits from this improved channel estimation if cancellation of inter-stream interference is applied. In link level simulation it was shown that its working point as function of SNR can be improved by more than 5dB for 10% block error rate.

Currently it is discussed in 3GPP to extend MIMO to 4 data streams to be sent in parallel (see, e.g., [11]). As this requires 4 pilot signals to be transmitted an even stronger mutual distortion impact of pilot signals due to scrambling can be expected. Although it is far from being clear how the

pilot scheme will be extended, scrambling induced deviation from orthogonality has to be included in the course of its definition.

ACKNOWLEDGMENT

The Author would like to thank Elke Hallmann and Klaus Knoche for many inspiring discussions during this study.

REFERENCES

[1] W. Aichmann, *Scrambling-Exact Channel Estimation in W-CDMA Systems*, in Proc. ICWMC, IARIA Luxembourg June 2011, p. 12-17. [Online] Available: http://www.thinkmind.org/index.php?view=article&articleid=icwmc_2011_1_30_20028

[2] S. McBeath, M. Ahmed, and K. Rohani, *Impact of imperfect estimators on W-CDMA receiver performance with MIMO antenna systems*, in Proc. IEEE 58th Vehicular Technology Conference (VTC 2003-Fall), vol. 2, Oct. 2003, pp. 1152-1156. [Online]. Available: http://ieeexplore.ieee.org/ie15/9004/28569/01285202.pdf?tp=&arnumber=1285202.

[3] Ch. Mehlführer, and M. Rupp, *Novel Tap-wise LMMSE Channel Estimation for MIMO W-CDMA*, in Proc. 51st Annual IEEE Globecom Conference, 2008, New Orleans, LA, USA, Nov. 2008. [Online]. Available: http://publik.tuwien.ac.at/files/PubDat_169129.pdf.

[4] B. Lindoff, Ch. Östberg, and H. Eriksson, *Channel estimation for the W-CDMA system, performance and robustness analysis from a terminal perspective*, in Proc. IEEE 49th Vehicular Technology Conference (VTC 1999), vol. 2, July 1999, pp. 1565-1569. [Online] Available: http://ieeexplore.ieee.org/iel5/6314/16887/00780649.pdf?tp=&arnumber=780649.

[5] I.R.S. Casella, E.S. Sousa, and P.J.E. Jeszensky, *Chip space-time block coding using iterative channel estimation with inter-antenna interference cancellation for W-CDMA systems using long scrambling codes*, in Proc. IEEE 15th International Symposium on Personal, Indoor and Mobile Rdio Communications, (PIMRC 2004), vol. 4, pp. 2313-2317. [Online] Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1368732&isnumber=29952.

[6] j.-W. Choi and Y.-H. Lee, *Evaluation of MIMO Transmission for HSUPA*, IEEE International Conference on Communications, 2003, vol. 1, pp. 748-752

[7] E. Dahlman, S. Parkvall, J. Sk̈old, and P. Beming, *3G Evolution, HSPA and LTE for Mobile Broadband*, 1st ed., Elsevier, 2007.

[8] 3GPP, *Technical specification group radio access network; Physical channels and mapping of transport channels onto physical channels (FDD) (Tech. Spec. 25.211 V8.4.0)*, 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/25_series/25.211/.

[9] L. Fathi, G. Jourdain, and M. Arndt, *A Bernoulli-Gaussian Model for Multipath Channel Estimation in Downlink WCDMA*, IEEE/SP 13th Workshop on Statistical Signal Processing, 2005, pp. 651-656

[10] H. Holma and A. Toskala, *HSDPA/HSUPA for UMTS*, John Wiley & Sons, 2006.

[11] White paper, *The Evolution of HSPA*, 4G Americas, 2011, [Online] Available: http://www.4gamericas.org/documents/4G\%20Americas\%20White\%20Paper_The\%20Evolution\%20of\%20HSPA_October\%202011x.pdf.

[12] 3GPP, *Technical specification group radio access network; Physical layer procedures (FDD) (Tech. Spec. 25.214 V8.5.0)*, 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/25_series/25.214/.

[13] For HSUPA, it is suggested in 3GPP to use the same pre-coding for pilot and data, which enables non-serving Node Bs to decode the data without signalling them the pre-coding index.

[14] M. Wrulich, S. Eder, I. Viering and M. Rupp, *Efficient Link-to System Level Model for MIMO HSDPA*, Proc. of the 4th IEEE Broadband Wireless Access Workshop, New Orleans, December, 2008.

[15] M. Wrulich, Ch. Mehlführer and M. Rupp, *Managing the Interference Structure of MIMO HSDPA: A Multi-User Interference Aware MMSE Receiver with Moderate Complexity*, IEEE Transactions on Wireless Communications, 2010, vol. 9, No. 4, pp. 1472-1482

[16] A. Szabo, *Evaluation of MIMO concepts in HSDPA by system level simulations*, Shaker Verlag, Aachen 2008. [Online]. Available: http://www.shaker.nl/Online-Gesamtkatalog-Download/2011.09.12-14.18.08-192.100.130.238-radB6669.tmp/3-8322-7822-2_INH.PDF

[17] 3GPP, *Technical specification group radio access network; Spreading and modulation (FDD) (Tech. Spec. 25.213 V9.1.0)*, 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/25_series/25.213/.

[18] J. G. Proakis and M. Saleh, *Digital Communications*, 5th ed., McGraw-Hill, 2009.

[19] ETSI Technical Report, *Universal Mobile Telecommunications System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS (UMTS 30.03 version 3.2.0)*, TR 101 112, V3.2.0 (1998-04). [Online]. Avalable: http://www.etsi.org/deliver/etsi_tr/101100_101199/101112/03.02.00_60/tr_101112v030200p.pdf

[20] Markus Reinhardt, *Analytical Performance Evaluation of Downlink Beamforming with Power Control by Transmit Power Limited Channel Inversion*, AEÜ Int. J. Electron. Commun., vol. 59, 2005, No. 2, pp. 1-7

Radiation Pattern Characteristics of Multiple Band-Notched Ultra Wideband Antenna with Asymmetry Slot Reconfiguration

Yusnita Rahayu
Department of Electrical Engineering
Faculty of Engineering, University of Riau
Pekanbaru 28293, Riau, Indonesia
E-mail: vannebula2001@yahoo.com

Razali Ngah
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
Johor Bahru, 81310, Malaysia
E-mail: razalin@fke.utm.my

***Abstract*- Radiation pattern characteristics of reconfigurable asymmetry slotted ultra wideband antenna with multiple band-notched is presented. The proposed antenna is planar monopole, consisting of a microstrip feed line and modified L and U slots with a pentagonal patch on the front side and half ground plane is on the back side of the board. The measurements were conducted by using Radio Frequency measurement and instrumentation facilities, software tools available at the Wireless Communication Center in Universiti Teknologi Malaysia. The modified antennas are designed for having reconfigurable frequency notched at Fixed Wireless Access, High Performance Local Area Network, and Wireless Local Area Network. A good agreement is achieved between simulated and measured Voltage Standing Wave Ratio results. The measured result is approximately less than 2.2 over 3 GHz to 10 GHz. It was also observed that the measured radiation patterns, H-planes, are omnidirectional with a slight gain decreased at boresight direction for measuring frequencies. There are also more ripples occurred in the measured pattern compared with the simulated one. Comparison measured radiation patterns between asymmetry slotted antenna with and without notch band characteristic have been done. Multiple reflections, rotator alignment, probe rotary joint and room scattering are found as primary sources of errors that cause degradation in measured radiation patterns. Current distribution characteristic of the antenna is also reviewed.**

Keywords- ***reconfigurable antenna, slotted antenna, radiation pattern, ultra wideband, band-notched antenna.***

I. INTRODUCTION

Reconfigurable antenna is defined as the antenna capability to reconfigure several parameters such as operating frequencies, impedance bandwidth, polarizations, and radiation patterns to fulfill operating requirements. The development of this antenna posses significant challenges to both antenna and system designers. The challenges lie not only in obtaining the desired level of antenna functionality but also in integrating this functionality into complete systems to arrive at efficient and cost effective solution.

This section reviews the concept of reconfigurable antennas that reuse their entire geometry for band-notched frequency Applications. Paper [1] is taken as a reference paper for this extension works. The concept is that each operating band resonates on the portion of the entire antenna geometry, therefore there is almost no extra size required to create multiband characteristics. A frequency shift can be achieved by connecting or disconnecting one or more antenna elements. Every configuration of an antenna will have a different resonating frequency and different radiation pattern. An antenna has to be designed with the objective that it should exhibit banding-notched behaviour and has to be compact. Various techniques used have been reported in literature [1][2][3][4][5][6][7][8][9][10][11][12][13][14][15].

In [2], there are two varieties of slotted antennas which have a frequency notched reported, a triangular notch and elliptical notch. Both antennas have frequency notch characteristics where the arc length of slots form a half wavelength resonance structure of particular frequencies, thus a destructive interference takes place causing the antennas to be non-responsive at that frequency. Other types of this kind antenna was reported in [5], a band-notched UWB (ultra wideband) antenna using a slot-type SRR (split ring resonator) was found very effective in rejecting unwanted frequency, such as that for WLAN (Wireless Local Area Network) service, in terms of its selectivity and small dimension. The SRR is composed of two concentric split ring slots and proposed for band stop application, since it provides a high Q characteristic. The slotted SRR was positioned near the feeding point to provide more coupling with the field. A multiple band-notched planar monopole antenna using multiple U-shape slots for multi band wireless system was also presented in [6]. The half wavelength U-shape slots were symmetrically inserted in the center of the planar element. In order to generate the two bands-notched characteristics, three U-shape slots were proposed.

The notch bands were also obtained by inserting an L-shaped slot in the ground plane [13]. The proposed antenna was fed by CPW (coplanar waveguide). By altering the length of the L-shaped slot, it can easily adjust the notched frequency. The similar techniques were also proposed in [14], the reject frequency was obtained by cutting a rectangular slot in the CPW ground which reduces the potential interference between UWB systems and the narrow systems. Thus, it is found that the use of slots for a patch or ground plane significantly provide band-notched characteristics.

An alternative antenna design without using the slot to obtain band-notched characteristic was proposed in [7]. The antenna consists of two same size monopoles and a small strip bar at the center showing the band rejection

performance in the desired frequency bands. The use of strip bar leads to high impedance at the notch frequency. The total length of strip bar from the ground plane is equal to a quarter-wavelength at the notch frequency.

In this paper, novel reconfigurable UWB antennas are designed by adopting the half wavelength slot structure technique. This technique has been found and proven in [2][5][6][13][14] as a very effective technique in providing band-notched characteristics. This paper mainly focuses on the reconfigurable notch band through the introduction of new slots, L and U slots, on the pentagonal patch antenna. This asymmetry couple slot, L and U, creates a new novelty slot design. Most published papers use a uniform and symmetry slot structure for their antenna design. In this paper, it is shown that the asymmetry slots can be designed to form band-notched characteristics with comparable results to others. This slot structure chosen is mostly influenced by current distribution surrounding the antenna patch.

Section II will discuss the antenna geometry and its current distribution characteristics. The techniques to design new reconfigurable slotted antennas having band-notched frequencies are also presented. The band-notched operation is achieved by incorporating some small gaps instead of pin diodes into the slot antenna. The term of small gaps in this paper will refer to switches. The switches are used to short the slot in pre-selected positions along the circumference. The length of the slot antenna can be lengthened or shorted by closing or opening the switches, allowing for a change in the notched frequency. Then the performances of reconfigurable antennas, in terms of the VSWR (voltage standing wave ratio) and radiation patterns, will be discussed and evaluated in Section III. Comparison measured radiation patterns between slotted antenna with and without band-notched characteristic is also discussed. Errors in measured radiation patterns are further analyzed and presented. Finally, a conclusion will be given in Section IV.

II. ANTENNA DESIGN

Fig. 1 shows the proposed antenna structure printed on the FR4 substrate of ε_r = 4.6. The pentagonal antenna is vertically installed above a ground plane (l_{grd}) of 11 mm. The optimum feed gap (h) to the ground plane is found to be 1.5 mm. The dimension of substrate is chosen to be 30 x 30 mm^2 (W_{sub} x L_{sub}) in this study. The antenna has a pentagonal patch with a width (w) of 15 mm and a length (l) of 12 mm. This shape is as a variety of rectangular shape with bevel techniques. This antenna geometry is similar to the antenna geometry presented in [1].

The slots proposed on patch effectively change its electrical length over a very wide bandwidth. Slot dimensions of the proposed antenna are listed in Table 1. The slot width is 0.5 mm in order to improve the bandwidth above 10 GHz. Asymmetry slot design, L and U, on this proposed antenna results novelty structure of the slotted UWB antenna, this is as one of contribution in this paper.

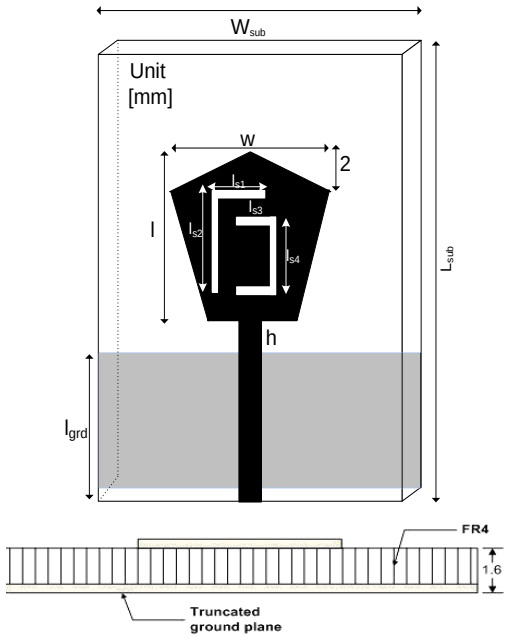


Figure 1. Geometry of L and U slotted antenna

TABLE I. DIMENSION OF L AND U SLOT

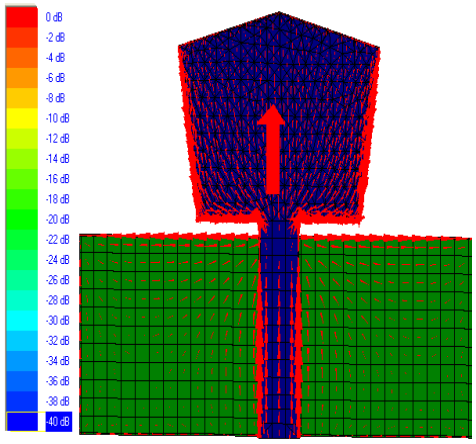
Description	L and U Slots	
	Symbol	*Size [mm]*
Slot length	Is1	6
	Is2	9
	Is3	3
	Is4	6.5

A. Current Distribution Characteristics

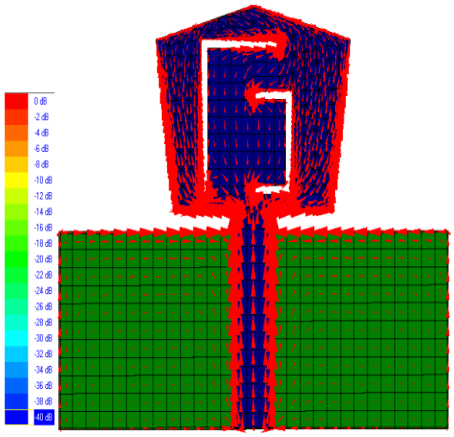
The geometry of the antenna implies the current courses and makes it possible to identify the active zone in the antenna. Active zone is the matching and radiator zone. Acting on matching and radiating areas allows controlling the bandwidth [12].

The couple slots, L and U, are designed very carefully by studying the current flow distribution which will give input impedance improvement. The study of the current distribution for bandwidth enhancement has been conducted in [16]. The current flows on a planar monopole antenna reveals that it is mostly concentrated in the vertical and horizontal edges as shown in Fig. 2(a). The current distribution on patch surface is disturbed by the introduction of new slots, L and U, as presented in Fig. 2(b).

It is observed that the horizontal current distributions are focused on the bottom edge of pentagonal patch. The discontinuity occurred from beveling at the bottom side of the patch and has enforced the excitation of vertical current mode in the structure, which presents a very wide bandwidth.



(a)



(b)

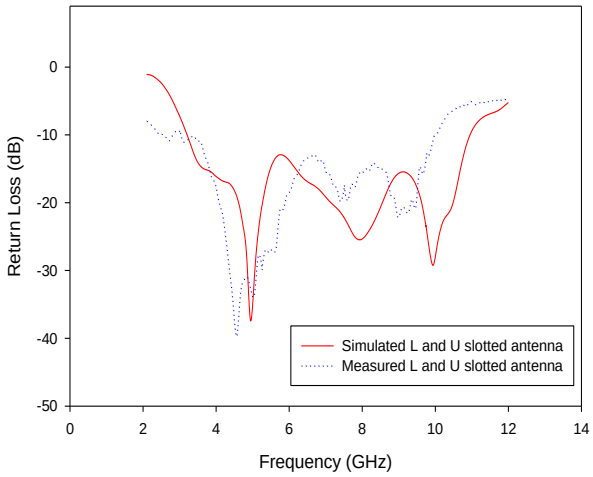
Figure 2. Current distribution characteristics

From Fig. 2(b), the current distribution is less of the area between both slots. However, these slots have resulted more vertical currents flowing to the patch. The U slot is designed to improve the upper dip resonance and the L slot improves the lower dip resonance. Individual L and U slot current distribution with their simulated results was reviewed in [17]. The coupling both slots has shown a good return loss with respect to -10 dB over UWB frequency range as presented in Fig. 3(a).

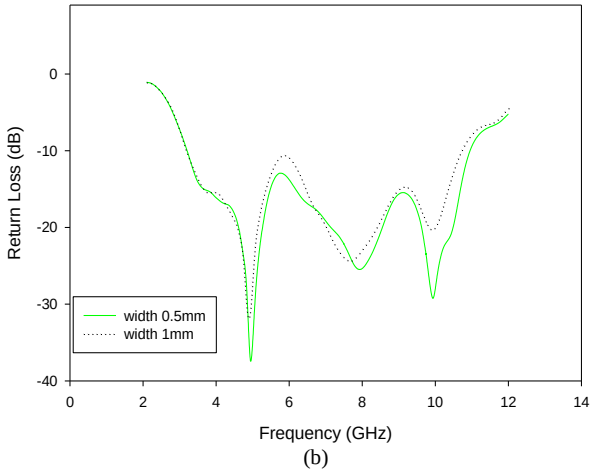
B. Simulated and Measured Return Loss

The simulated and measured return loss is presented in Fig. 3(a). The measured return loss is slightly shifted to the simulated one, but they still cover 2.5 GHz to 10.1 GHz as what the UWB required. The length of L slot is 14.5 mm approximately equal to 0.25λ at 5.3 GHz, and the length of U slot is 11.5 mm approximately equal to 0.4λ at 10.3 GHz.

It is also observed that the width of slots chosen affect to the return loss of dip resonance especially at an upper frequency as shown in Fig. 3(b). The width of 0.5 mm results -28 dB of return loss at 10.3 GHz.



(a)



(b)

Figure 3. Simulated and measured return loss

This measurement was done by using a 50-Ω SMA connector which is soldered at the bottom edge of microstrip line and connected to the network analyzer by an RF (Radio Frequency) cable. The RF cable significantly affects the performance of the AUT (antenna under test). However, some differences in the simulated and measured results are expected, since in the simulation model the mismatch due to the adapter and connector used are not taken into consideration. In reality the coaxial cable has a considerable effect, especially the length of its inner conductor, which is connected to the input of the antenna, creating an additional inductance. In addition, since the antenna is fed by a microstrip line, misalignment can result because etching is required on both sides of the dielectric substrate. The alignment error results degradation to the antenna performance.

C. Reconfigurable Techniques

As mentioned in Section 1, the half wavelength slot structure technique is used in designing the proposed UWB slotted antenna. The main purpose here is how to design the reconfigurable UWB slotted antenna without major modification from the previous antenna design presented in [16][17]. The unique of this proposed reconfigurable UWB antenna is the antenna has additional ability in rejecting the FWA (Fixed Wireless Access) band, whereas most existing papers focusing in rejecting WLAN and HIPERLAN (High Performance Local Area Network) bands. Licensed band at FWA for point to multipoint radio systems assigned by MCMC (Malaysian Communications and Multimedia Commissions) for 3.4 to 3.7 GHz is considered giving potential interference to UWB application. This is due to the allocation frequency for this FWA within the UWB range. Therefore, recently the consideration of UWB antennas is not only focused on an extremely wide frequency bandwidth, but on the ability of rejecting the interference from WLAN 11.a (5725–5825 MHz) and HIPERLAN (5150–5350 MHz) within the same propagation environment. Thus, the proposed notched antenna is not only designed to reject interference from WLAN, HIPERLAN but also from FWA.

The basic geometry of L and U slotted antenna will be taken as a reference to form a new modified L and U reconfigurable slotted antennas. There are maximum six switches used to provide the reconfigurable function. No especial matching network is used and the matching properties are solely determined by the placement of the switches.. The number of switches used and their placement are based on the notched band required. The on-off switches will determine the total slot length at a certain notched band. With the notched band characteristic, the antenna allows to reconfigurable its frequency that only responds to other frequencies beyond the rejection bands within the UWB bandwidth. However, the dimensions of the antenna and substrate are kept equal to the original model.

For the simulation purposed, the switches are considered as ideal switches and are modeled as small patches that connect or disconnect the adjacent slot, changing the antenna's slot length.

In Fig. 4, the switches have different colors for on and off state condition. Blue color represents the on state condition and red color for the off state condition. In order to provide the UWB characteristic, the switches are placed as shown in Fig. 4(a). Three switches of #2, #3, and #4 are in the off state position. Other switches of #1, #5, and #6 are in the on state condition. When the switches are in the off state condition, the gap between slots occurs and the current flowing to the gap. When the switches are in the on state condition, there is no current flowing to the slots. Thus it forms continuous slots. The switches of #2 and #3 are incorporated to the first additional slot (I_{s20}) which is 3.5 mm of slot length. The switch of #4 is attached to the second additional slot (I_{s21}) which is 2.5 mm of slot length. The two additional slot lengths, I_{s20} and I_{s21}, are very critically determined by the frequency notched band characteristics.

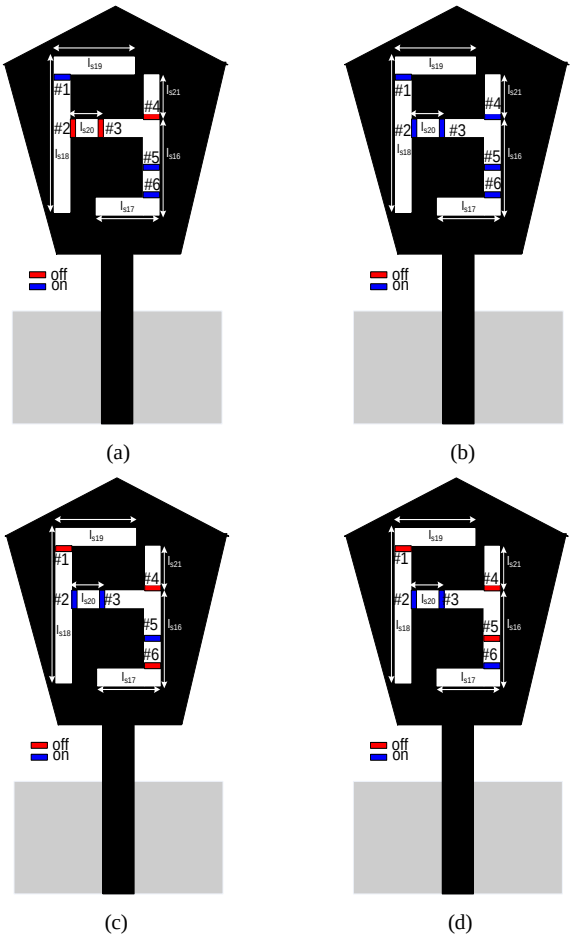


Figure 4. Switching configuration for L and U slotted antennas: (a) without notched, (b) notched at FWA, (c) notched at HIPERLAN, (d) notched at WLAN

To accommodate the notch frequency at FWA band, the length of U slot is lengthened than the previous model. Thus, six switches are used and placed systematically. The frequency notched characteristic antenna at FWA is shown in Fig. 4(b). All switches are in the on state position (continuous slot). Total slot lengths are 32 mm or approximately equal to 0.4λ at 3.7 GHz. The total slot lengths mean the sum of slot lengths of L, U and additional slots. Fig. 4(c) and Fig. 4(d) present the frequency notched characteristic antenna at HIPERLAN and WLAN, respectively. To reject interference from HIPERLAN, the switches of #1, #4, and #6 are in the off state position while switches of #2, #3, and #5 are in the on state position. It is investigated that by inserting those switches in the off state condition broke the connection between slots. This break connection has reduced the slot length to be 20.75 mm or approximately equal to 0.33λ at 5.2 GHz. Thus, the antenna has a frequency notched at HIPERLAN. The total slot length is measured by the length of connecting slots.

The configuration of switches in Fig. 4(d) have resulted an antenna with frequency notched at WLAN. It is shown that the switch of #5 is set in the off state position in order to reduce the slot length, while the switch of #6 is set in the on

state position. This is the only difference while comparing to the HIPERLAN configuration. Total slot lengths are 18 mm or approximately equal to 0.33λ at 5.75 GHz and measured by the length of connecting slots.

For prototype development, the gaps are created in the UWB antenna pattern, which are represented as switches. The dimensions of switches used in these prototypes are 0.7 mm× 1 mm. Although this model is ideal, it gives a good approximation for the commercial pin diode switches. The photograph of the antenna prototype that has a band-notched at FWA is shown in Fig. 5.

III. EXPERIMENTAL VERIFICATION

Fig. 6 shows the simulated and measured VSWR for reconfigurable modified L and U slotted antennas. By varying the slot lengths and break the connection between slots using switches, the proposed frequency notched is achieved.

The simulated results show the FWA notched band obtained from 3.57 GHz to 3.86 GHz with the total slot length of 32 mm at 3.7 GHz, which is the center frequency. While the HIPERLAN and WLAN notched bands are from 4.84 GHz to 5.33 GHz and 5.53 GHz to 6.02 GHz, respectively. It is noted that beyond the frequency notched bands, the VSWR is kept to be less than 2.

The measured VSWR is approximately less than 2.2 over 3 GHz to 10 GHz, except for the antenna with band-notched at HIPERLAN band. The VSWR beyond 9.5 GHz is getting worse. For band-notched at FWA, the rejection is in the range of 3.76 GHz to 4.25 GHz, and the rejection of HIPERLAN is 5.23 GHz to 5.53 GHz. The antenna with band-notched at WLAN band has the rejection in the range of 5.72 GHz to 6.1 GHz. Even though the band-notched slightly shifted from the required bands, but it still covers the rejection bandwidth.

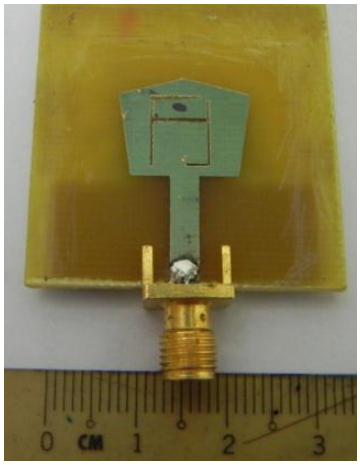


Figure 5. Photograph of FWA notch band antenna

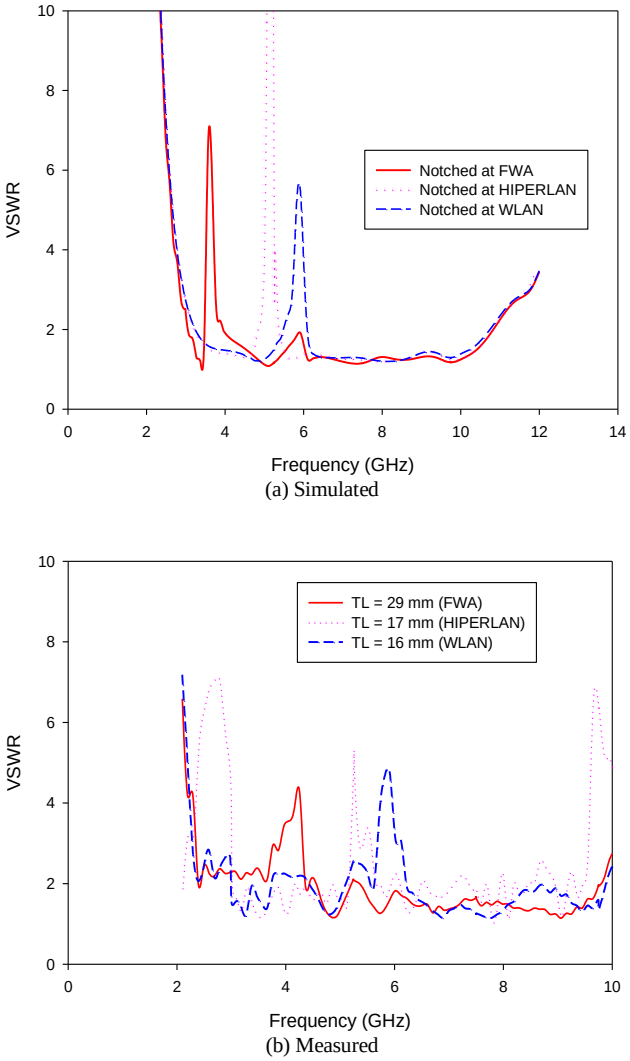


Figure 6. Simulated and measured VSWR for reconfigurable modified L and U slotted antennas

A. *Spherical Near-Field Testing*

Once the resonance frequencies were identified, principal radiation patterns were taken to characterize the operational performance of each antenna. These measurements were obtained using the indoor anechoic chamber room. For these measurements, the chamber was arranged as shown in Fig. 7.

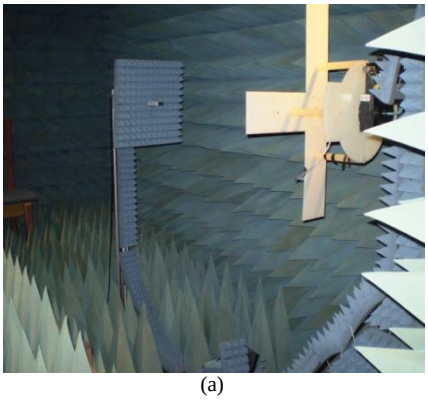
From Fig. 7, it shows the measurement setup for both antenna and probe which were mounted to the vertical positioned-holders. The probes available at the chamber are in the frequency ranges of 3.95 – 5.85 GHz and 8.95 – 12 GHz, respectively.

The existing chamber employed the spherical near-field measurement. By definition, near-field tests are done by sampling the field very close to the antenna on a known surface. From the phase and amplitude data collected, the

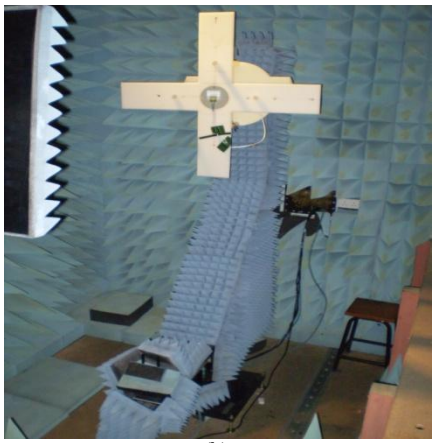
far-field pattern must be computed in much the same fashion that theoretical patterns are computed from theoretical field distributions. The transformation used in the computation depends on the shape of the surface over which the measurements are taken with the scanning probe [18].

This spherical system acquired a double data set on the AUT. The '360phi' data set is taken with full 360° phi rotation of the AUT, but with only 0° – 180° motion in theta. In this mode, the AUT's Z axis will only be looking at one side of the chamber during the measurement [19]. The AUT was swept every 2° increment in the azimuth plane in order to reduce the aliasing errors. Data point spacing aliasing errors will occur for spherical near-field measurements if the data point spacing is not small enough to sample the highest spatial frequency components in the measured data [19].

The radiation patterns were measured at 4 GHz and 5.8 GHz. The measured radiation patterns were plotted into H (horizontal) and V (vertical) cuts. The H-cut is cut for the azimuth plane with the fixed elevation angle at 0^0 and vary the azimuth angle. The V-cut is cut for the elevation plane with a fixed azimuth angle at 0^0 and vary the elevation angle.



(a)

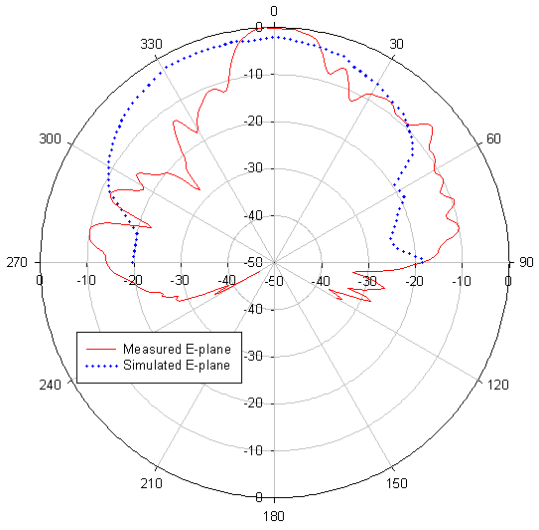


(b)

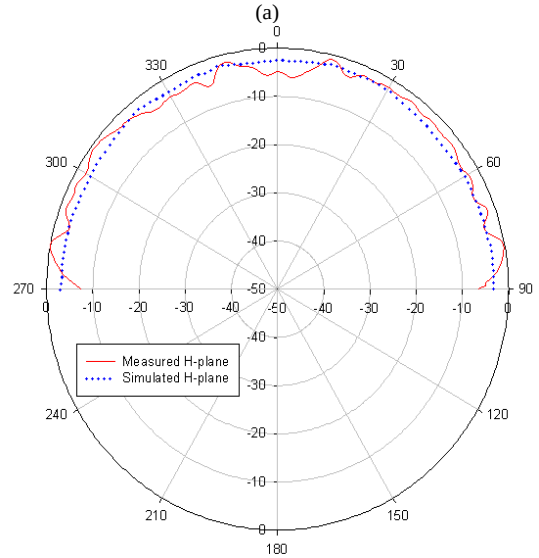
Figure 7. The radiation pattern measurement setup inside the anechoic chamber room

B. Radiation Pattern Measurement

Radiation pattern's measurements have been done for both proposed antennas with and without band-notched characteristic. Fig. 8 and Fig. 9 show the radiation pattern results at 4 GHz and 5.8 GHz for antenna without band-notched characteristic, respectively.



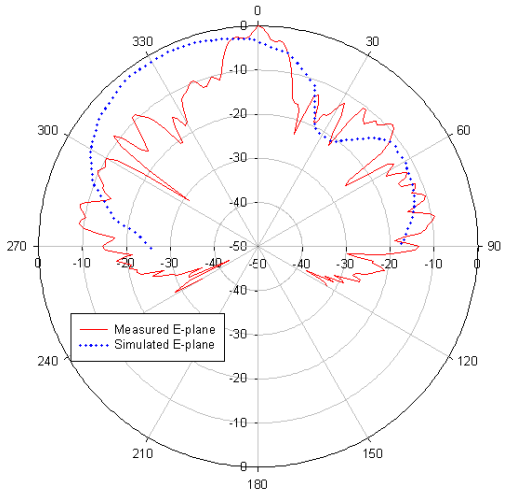
(a)



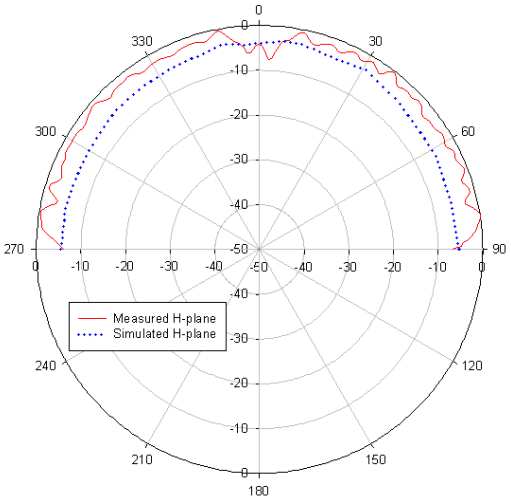
(b)

Figure 8. Simulated and measured E and H-planes for the slotted antenna at 4 GHz (without notched band)

All radiation patterns were measured and examined for both E and H-planes. They show acceptable results where the overall H-planes for both frequencies providing omnidirectional patterns. But, the more ripples occur in the measured E-planes radiation pattern. This is due to some errors during the measurement process.



(a)

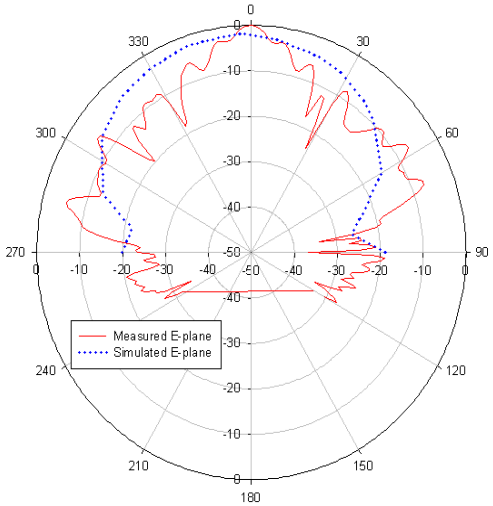


(b)

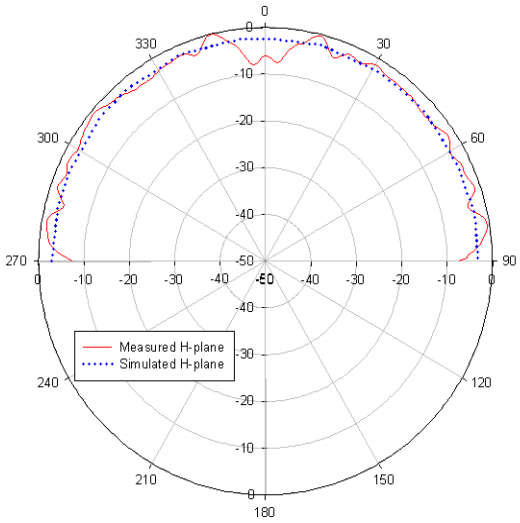
Figure 9. Simulated and measured E and H-planes for the slotted antenna at 5.8 GHz (without notched band)

Measured and simulated radiation patterns of reconfigurable antennas with band-notched characteristic are plotted in Fig. 10 to Fig. 15. Comparison radiation patterns for both proposed antennas have been done. From the results obtained, all H-plane patterns show nearly similar patterns, omnidirectional, for all measured frequencies. But, the E-plane patterns show the variation patterns for each frequency.

Fig. 10 and Fig. 11 show the radiation patterns of antenna notched at FWA for 4 GHz and 5.8 GHz, respectively. From both figures, there are slightly back-lobes present for the E-planes. Both H-planes are omnidirectional with a slight gain decreased at boresight direction. The H-planes patterns retain a satisfactory omni-directionality (less than 10 dB gain variation in most directions) over the entire bandwidth in both simulation and experimental.



(a)



(b)

Figure 10. Simulated and measured E and H-planes at 4 GHz for antenna notched at FWA

The E-plane pattern is the radiation pattern measured in a plane containing feed, and the H-plane pattern is the radiation pattern in a plane orthogonal to the E-plane.

The results show that the radiation patterns are changing as the frequency increases. The patterns resulted from the measurements have many ripples in amplitude due to many reflections into the field between the AUT and probe. The reflections may come from the room (floor and ceiling), chamber scattering, antenna holder itself and track inside the anechoic chamber. It is shown in previous Fig. 7, the antenna size is very small compared to its huge holder, and also the floor and the track surrounding the antenna tower are not all covered by absorber.

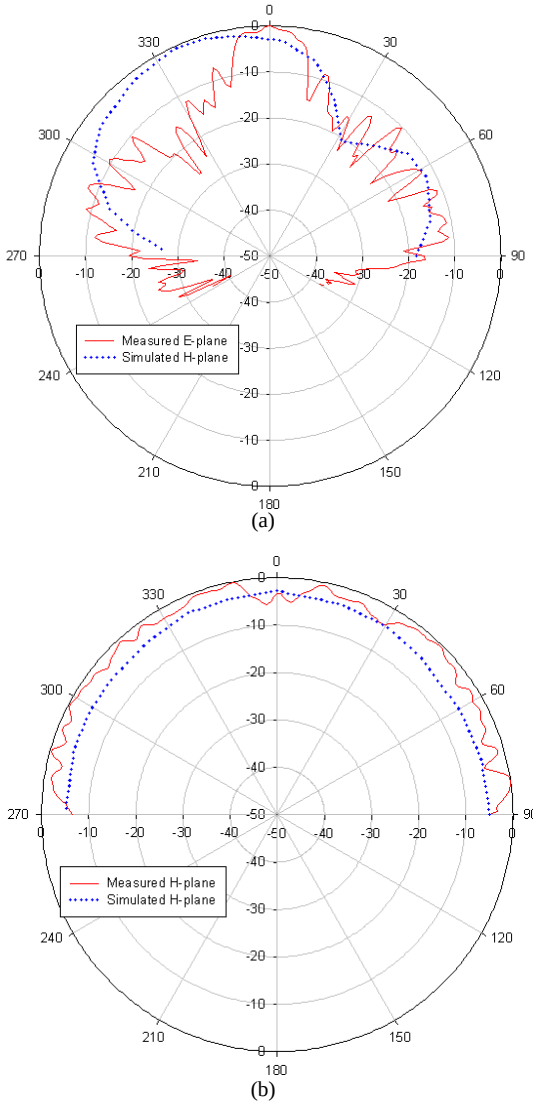


Figure 11. Simulated and measured E and H-planes at 5.8 GHz for antenna notched at FWA

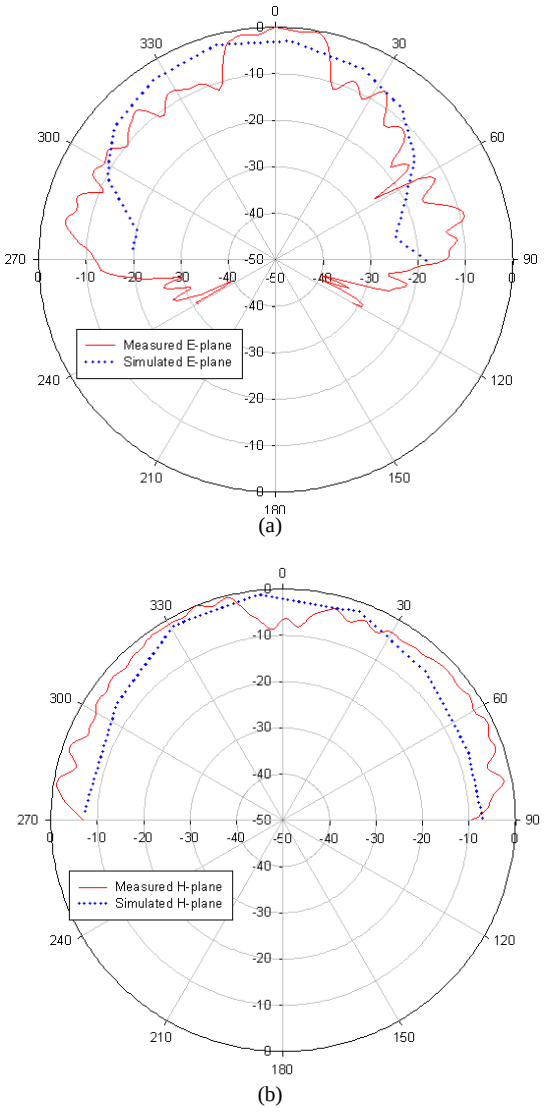
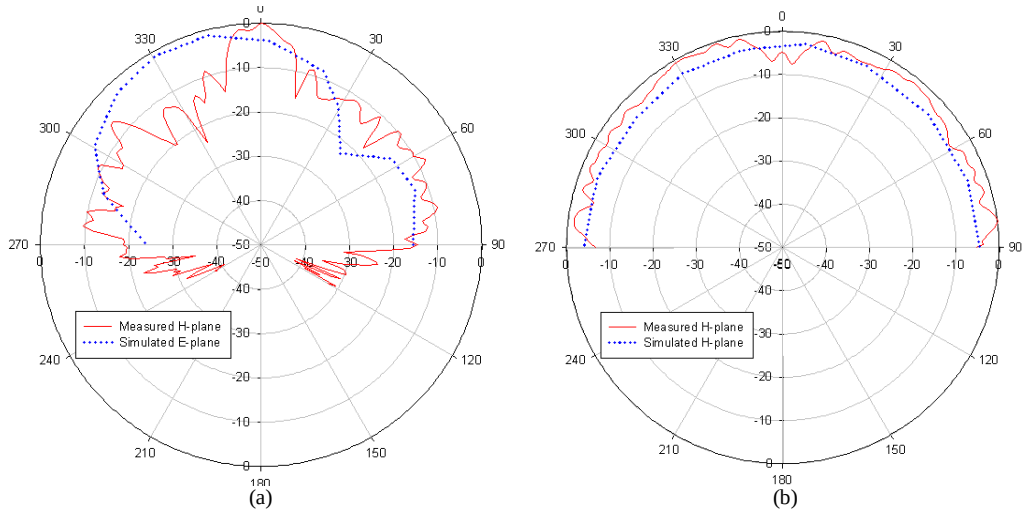


Figure 12. Simulated and measured E and H-planes at 4 GHz for antenna notched at HIPERLAN

Fig. 12 to Fig. 15 shows the radiation patterns for antenna notched at HIPERLAN and WLAN. The E and H-planes are measured at 4 GHz and 5.8 GHz for both testing.

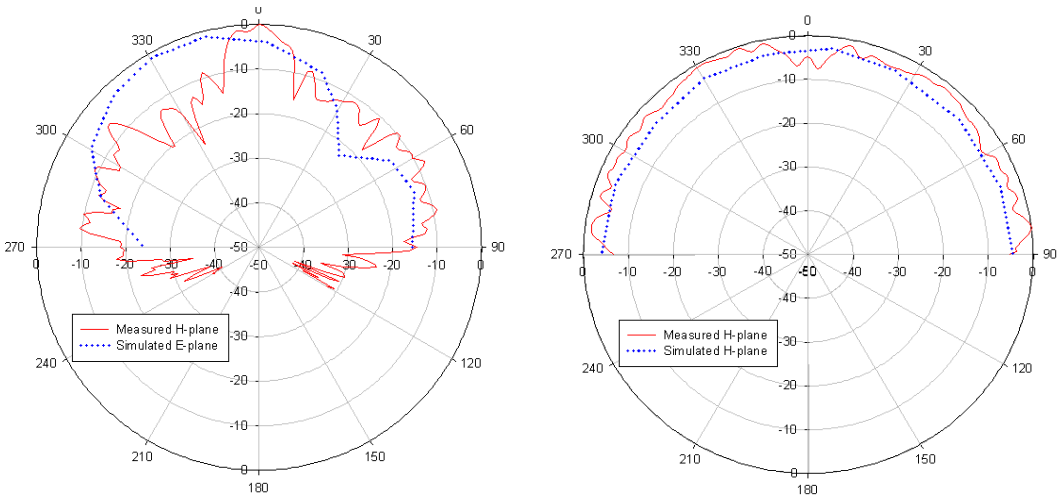
It is observed that the measured E-planes for antenna notched at HIPERLAN broader than the measured E-planes for antenna notched at FWA. However, the patterns resulted from the measurement are still having many ripples in amplitude. Due to the track surrounding the antenna tower are not all covered by absorber, if moved sideways about half cycle the ripples occur and if moved vertically the ceiling and floor reflections are indicated. The most significant is probably from improper cable connectors allowing excitation of the outside surface [18]. Leakage will be added to the measured pattern as degradation. Both H-planes are omnidirectional.

The measured radiation patterns demonstrate that at the low end of the operating band, the currents are well distributed over the patch antenna plate so the patterns are relatively broad. But at the high end of the operating band, the currents are concentrated near the slot, thus the fields radiate mainly through the slot.



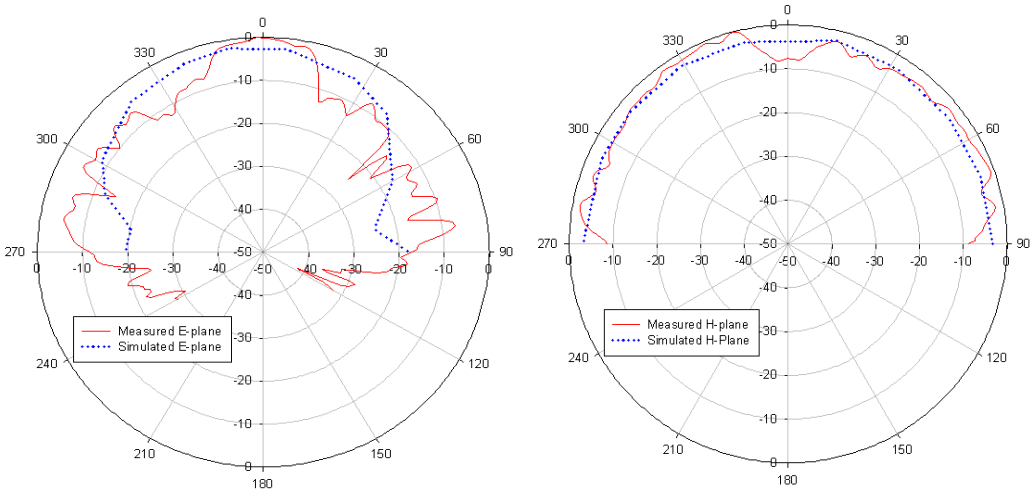
(a) (b)

Figure 13. Simulated and measured E and H-planes at 5.8 GHz for antenna notched at HIPERLAN



(a) (b)

Figure 14. Simulated and measured E and H-planes at 4 GHz for antenna notched at WLAN



(a) (b)

Figure 15. Simulated and measured E and H-planes at 5.8 GHz for antenna notched at WLAN

C. *Estimating Error Analysis in Radiation Pattern Measurement*

Since the measured radiation patterns obtained have many ripples and dips for the E-planes, the error analysis have been done to estimate the uncertainty and the caused of the error. The development of analysis and measurements to estimate the error during measurement in spherical near-field has been conducted in [19] [20].

During measurement process, several requirements are needed to take into consideration. It is observed that multiple reflections, rotator alignment, probe rotary joint and room scattering are shown as primary sources of errors that cause degradation in measured radiation pattern. Multiple reflections between AUT and probe produce the ripple in the measured data. The correction is difficult to do due to the need a series of near-field measurement at Z-positions separated by λ/8. The far-fields are then calculated for each and averaged [21]. In the case of multiple reflections and random errors, multiple measurements are required.

Fig. 16 shows the example of the random errors presented during measurement. These measured patterns are from repeated measurement at 5.8 GHz for L and U slotted antenna. Some corrections on the AUT alignments have also been done. From the pattern comparison graphics show the random error occurs in amplitude and phase during multiple measurements.

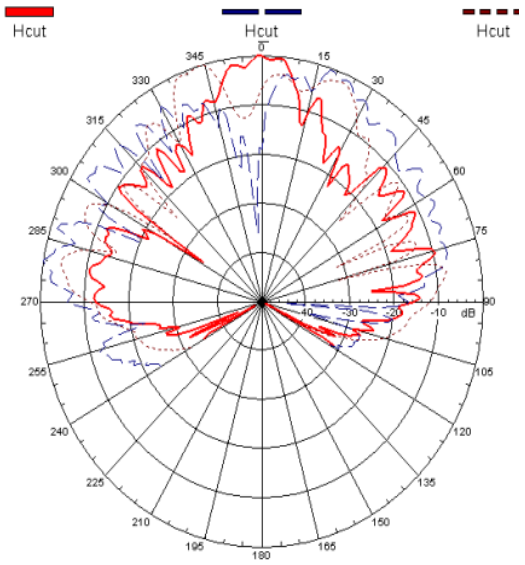


Figure 16. An example of results of random errors for L and U slotted antenna at 5.8 GHz

When the AUT is not precisely aligned to reference coordinate system, vector components or coordinate angles may change for some rotation. This correction can be done by realigning the AUT position.

The rotator alignment for spherical is a special case of position errors. The orthogonally and intersection of the theta and phi axes and the coincidence of the phi and probe polarization axes can be checked by measuring and comparing near-field cuts at θ = 0 and 180 degrees. The misalignment can be corrected by adjustments of the mechanical system. However, the difference values still present at those near-field cuts of measured data.

The probe rotary joints associated with the theta and phi rotators produce some amplitude and phase variations as it is moved and this is as a function of theta and phi. The error can be observed from the measured data by comparing the amplitudes and phases of the two components at (θ, φ) coordinates, (0, 0) and (0,180) [20]. At these points, the amplitude should be identical and the phases should be either identical or 180 degrees different. This error is more important at high frequencies where rotary joints may not be as accurate. From the measured data obtained, it is observed that the variations have presented in amplitudes and phases for those points of both proposed antennas. Thus, this error contributes to the uncertainty in far-field results.

Both probe and room scattering can produce higher spatial frequencies in the data. The smaller data point spacing is needed to avoid the aliasing errors. The room scattering effect can be more severe when low gain AUT's are being measured [20]. In this measurement, the 2^{o} data point spacing is used for scanner movement. Scattering from structures and absorber introduces an error that is difficult to estimate. This is because the procedure is demanding and time consuming. The AUT and probe must be translated together in a combination of X, Y and Z movements while maintaining precise angular alignment. The translations should be at least multiple wavelengths in dimension and the AUT must be realigned in the new position. Comparison of the patterns from the two locations provides an estimate of the room scattering but it is difficult to distinguish from alignment differences, probe/AUT multiple reflections and system drift [20].

It is concluded that obtaining true patterns depends primarily on accurately positioning the probe, accurately measuring the field, and eliminating distortions in the field introduced by the room, tracks, or probe. The room reflections must be lower than the basic side-lobes level and the probe must have low reflections. The probe position must be accurate to give better tolerance corresponding to the side-lobe level. In a spherical near-field range, the spherical measurement surface will be imperfect due to inaccuracies of the positioners and misalignment of these positioners [19].

IV. CONCLUSION

Novel reconfigurable asymmetry slotted antennas with band-notched characteristic at FWA, HIPERLAN, and WLAN band have been successfully designed and developed. The measured VSWR less than 2.2 is achieved. The radiation pattern behaviour for both proposed antennas with and without band-notched characteristic has been examined. The omnidirectional patterns are obtained for all measured H-plane patterns at 4 GHz and 5.8 GHz for both antennas. The E-plane patterns show a unique pattern for each frequency. It tends to radiate broadly in all directions but more ripples occur due to some errors during the

measurement process. Current distribution characteristic for the asymmetry slotted antenna is presented as well. It is shown that the introduction of asymmetry slots has disturbed the current flow on pentagonal patch. Six switches are attached to the antenna for reconfigurable purpose. By varying the slot length, the frequency notched antennas are performed at certain frequencies. Errors in measured radiation patterns are further analyzed. It is concluded that several requirements need to take into consideration during measurement process such as rotator alignment, room scattering and probe rotary joint.

ACKNOWLEDGEMENT

The authors would like to thank the Wireless Communication Center (WCC) of Universiti Teknologi Malaysia for their supporting and facilitating this research.

REFERENCES

[1] Yusnita Rahayu, Azlyna Senawi, and Razali Ngah, "Radiation Pattern Behaviour of Reconfigurable Asymmetry Slotted Ultra Wideband Antenna," Proc. The Seventh International Conference on Wireless and Mobile Communications (ICWMC 2011), June. 2011, pp. 81-86, ISBN: 978-1-61208-140-3.

[2] H. G. Schant, G. Wolenec, and E. M. Myszka, "Frequency Notched UWB Antenna," Proc. IEEE Conference on Ultra Wideband Systems and Technologies (UWBST), Nov. 2003, pp. 214-218, doi: 19.1109/UWBST.2003.1267835.

[3] Yongjin Kim and Do-Hoon Kwon, "CPW-Fed Planar Ultra-Wideband Antenna Having a Frequency Band Notch Function," Electronics Letters, vol. 40(7), April. 2004, pp. 403-405, doi: 10.1049/el:20040302.

[4] Saou-Wen Su, Kin-Lu Wong, and Fa-Shian Chang, "Compact Printed Ultra-Wideband Slot Antenna with a Band Notched Operation," Microwave & Optical Technology Letters, vol. 45(2), April. 2005, pp. 128-130, doi: 10.1002/mop.20746.

[5] J. Kim, C. S. Cho, and J. W. Lee, "5.2 GHz Notched Ultra-Wideband Antenna Using Slot-type SRR," Electronics Letters, vol. 42(6), March. 2007, pp. 315-316, doi: 10.1049/el:20063713.

[6] Wang-Sang Lee, W. G. Lim, and Jong-Won Yu, "Multiple Band-Notched Planar Monopole Antenna for Multiband Wireless Systems," IEEE Microwave and Wireless Components Letters, vol. 15(9), Sept. 2005, pp. 576-578, doi: 10.1109/LMWC.2005.855374.

[7] Seokjin Hong, Kyungho Chung, and Jaehoon Choi, "Design of a Band-Notched Wideband Antenna for UWB Applications," Proc. IEEE Antennas and Propagation Society International Symposium, July. 2006, pp. 1709-1712, doi: 10.1109/APS.2006.1710892.

[8] K. C. L. Chan and Yi Huang, "A Compact Semi-Circular Disk Dipole with Notched Band for UWB Applications," Proc. The Institution of Engineering and Technology Seminar on Ultra Wideband Systems, Technologies, and Applications, April. 2006, pp. 226-230, ISBN: 0-86341-625-X.

[9] Kyungho Chung, Jaemoung Kim, and Jaehoon Choi, "Wideband Microstrip-Fed Monopole Antenna Having Frequency Band Notch Function," IEEE Microwave and Wireless Components Letters, vol. 15(11), Nov. 2005, pp. 766-768, doi: 10.1109/LMWC.2005.858969.

[10] Yongjin Kim and Do-Hoon Kwon, "Planar Ultra Wideband Slot Antenna with Frequency Band Notch Function," Proc. IEEE Antennas and Propagation Society International Symposium, vol. 2, June. 2004, pp. 1788-1791, doi: 10.1109/APS.2004.1330545.

[11] Tamer Aboufoul and Akram Alomainy, "Reconfigurable Printed UWB Circular Disc Monopole Antenna," Proc. IEEE Loughborough Antennas & Propagation Conference (LAPC), Nov. 2011, pp. 1-4, doi: 10.1109/LAPC.2011.6114078.

[12] I. Pele, Y. Mahe, A. Chousseaud, S. Toutain, and P. Y. Garel, "Antenna Design with Control of Radiation Pattern and Frequency Bandwidth," Proc. IEEE Antennas and Propagation Society International Symposium, vol. 1, June. 2004, pp. 783-786, doi: 10.1109/APS.2004.1329787.

[13] Luyang Ji, Guang Fu, Jiayue Zhao, and Qiyilu, "CPW-fed UWB Antenna with the Design of Controlable Band Notch," Proc. IEEE International Conference on Electronics, Communications and Control (ICECC), Sept. 2011, pp. 788 – 790, doi: 10.1109/ICECC.2011.6067554.

[14] Henan Wang and Ying Song Li, "Bandwidth Enhancement of A Wide Slot UWB Antenna with A Notch Band Characteristic," Proc. IEEE 3rd International Conference on Communication Software and Networks (ICCSN), May. 2011, pp. 365-368, doi: 10.1109/ICCSN.2011.6014070.

[15] R. Movahedinia, M. Ojaroudi, and S. S. Madani, "Small Modified Monopole Antenna for Ultra-Wideband Application with Desired Frequency Band-Notch Function," IET Microwaves, Antennas & Propagation, Aug. 2011, vol. 5(11), pp. 1380-1385, doi: 10.1049/iet-map.2009.0560.

[16] Yusnita Rahayu, Tharek Abd. Rahman, Razali Ngah, P. S. Hall, Numerical Analysis of Small Slotted Ultra Wideband Antenna Based on Current Distribution for Bandwidth Enhancement, Ultra-Wideband, Short Pulse Electromagnetics 9 (UWB SP 9), 2010, Springer Publication, Part 4, pp. 215-223, doi: 10.1007/978-0-387-77845-7_25.

[17] Yusnita Rahayu, Tharek Abd. Rahman, Razali Ngah, P. S. Hall, "Small Printed Ultra Wideband Antenna with Having A Coupled Slot," Proc. IEEE Loughborough Antennas & Propagation Conference (LAPC), Nov. 2009, pp. 397-400, doi: 10.1109/LAPC.2009.5352426.

[18] Gary E. Evans, Antenna Measurement Techniques, Boston, London: Artech House. 1990.

[19] G. Hindman and Allen C. Newell, "Simplified Spherical Near-Field Accuracy Assessment," Technical Paper, Nearfield System, Inc. 2004.

[20] Allen Newell and Greg Hindman, "Planar and Spherical Near-Field Range Comparison with -60 dB Residual Error Level," Technical Paper, Nearfield System, Inc. 2004.

[21] E. Antonino-Daviu, M. Cabedo-Fabres, M. Ferrando-Bataller, and A. Valero-Nogueira, "Wideband Double-Fed Planar Monopole Antennas," Electronics Letter, vol. 39(23), Nov. 2003, pp. 1635–1636, doi: 10.1049/el:20031087.

Effect of Global and Local Brightness on Quality of 3D and 2D Visual Perception

Mahsa T. Pourazad
TELUS Communications Company, Canada
University of British Columbia, Canada
pourazad@ece.ubc.ca

Zicong Mai, Panos Nasiopoulos
University of British Columbia, Canada
{zicongm, panos}@ece.ubc.ca

Abstract—**There are several different factors that affect the perceived quality of 3D content. Our objective in this paper is to investigate how the global and local change of brightness in a scene will affect the overall 3D visual perception. Our work contains two studies: 1) global change of brightness, which is achieved by setting different exposure times on the camera, and 2) local change of brightness, where the brightness of the background is consistent and the object's brightness is changed using an external light source and/or a reflector disk. Subjective evaluations were performed, with the subjects being asked to rate the 3D perceptual quality of each sequence. The results of both studies demonstrate marginal difference in the brightness effects between 3D and 2D videos. The latter study showed that the Weber contrast between objects of interest and background should be within the range of -0.35 to 0.55 to provide viewers with high quality 3D experience.**

Keywords-3D TV; brightness; quality of experience; 3D perception; contrast

I. INTRODUCTION

Recently, 3D video has received increased attention among investors, researchers and technology developers. The introduction of 3D TV can only be a lasting success if the perceived image quality provides a significant step up from the conventional 2D television, while maintaining the same viewing comfort. The availability of high quality 3D content will also be a key factor to this success. Recoding 3D content – let alone high quality - is much more demanding and challenging than 2D video. In general, 3D content production needs different considerations and provisions beside the ones found in the conventional 2D video production to guarantee high quality of the produced content and the comfort of viewers.

There are many factors and parameters that could affect the perceptual quality of 3D media. While the effects of different acquisition parameters on 3D perception have been studied before, their influence on the perceived quality has not been quantitatively assessed. More research and studies are required in order to improve our understanding of different factors that affect the viewers' perception of 3D video content. This knowledge will allow us to capture high quality 3D content that may help reduce or even eliminate the visual discomfort of the viewers and thus improve the overall quality of experience. To this end, the study by Pourazad et al. (CONTENT 2011) quantitatively investigates the effect of contrast (local brightness-change) on the visual quality of 3D content [1]. Scene detail has also been noted to affect the visual comfort and quality of 3D content [1]. Goldmann et al. has addressed the effect that the distance between stereo cameras has on the perceptual quality of the captured videos [3]. The relation between the distance of the object(s) of interest from the camera and the quality of the perceived images when watched on different size displays has been investigated by Xu et al. [4].

Content producers recognize contrast and brightness as some of the important factors in capturing 3D content and based on their practical experience they suggest using higher light settings for capturing 3D videos compared to 2D ones [5]. In physiological studies it has been found that there is a relationship between contrast and the human visual depth perception [6]. In this paper, we perform a systematic study on the effect of the global and local brightness-change on the visual quality of 3D content, and compare it to that of 2D. This study is the extension of our previous work on the effect of local brightness change (contrast) on 3D visual perception [1].

The objective of this study is two fold: first to provide guidelines for adjusting brightness levels during 3D capturing in order to offer the best 3D quality of experience; second to determine how these levels differ from the case of 2D video and use this knowledge as a guideline for capturing 3D content. In our study we investigate the effect of both global and local brightness-change on the perceptual quality of 3D content. We first examine the scenario where the brightness of the whole scene is altered (change of global brightness). Global brightness-change can be achieved by setting different exposure times on the camera. In the second part of our work only the brightness of the object(s) of interest is altered and the brightness of the rest of the scene stays relatively unchanged (change of local brightness). This can be achieved using a light source or reflector disks. Since there is no standardized set of 3D or 2D video clips that record the same (sufficiently similar) scene in different brightness/contrast levels, for both of these scenarios, we decided to capture outdoor and indoor scenes for our test. Note that synthetic videos are not practical in our study. Simply changing the brightness or contrast of a recorded video is a lossy process and the resulting videos may have unwanted artifacts, which will hamper the 3D perception. Once the test videos are captured we perform extensive subjective quality assessment experiments to quantify the

perceived quality of the 3D experience at different levels of global and local brightness. Understanding both of these scenarios will allow us to thoroughly identify criteria related to brightness for enhancing the viewers' 3D quality of experience.

The rest of the paper is structured as follows: Section II describes the experimental setup and discusses the results of the global brightness-change experiment, Section III provides the details of our subjective evaluations for the local brightness-change case, and conclusions are drawn in Section IV.

II. Experiment One – Global Change Of Brightness

In this section, we investigate the effect of global brightness on the perceptual quality of 3D content and compare it with that of 2D video. In this case, the captured brightness of the background and that of the object are changed simultaneously. The following subsections elaborate on the details of our experiment.

A. Capturing Setup

To capture test video sequences we use two identical full HD cameras (Sony HDR-XR500V 1080 60i NTSC) with baseline distance of 9cm. Fig. 1 shows the stereo camera setup used in our experiment. We used the same settings on both cameras, which were aligned in parallel and attached to a bar that was custom-made for this purpose. Subsequently, the bar was secured to a tripod. Since zoom lenses may differ [5], only the extreme ends of the zoom range were used to avoid any zoom correction post-processing. We set the cameras to manual mode with ISO equal to 100 and we disabled face-detection, and the backlight compensation mode. Note that the camera setup and configuration is more complex in the case of 3D video recording and stereoscopic geometry and camera calibration require special attention [7].

Figure 1. Stereo camera setup consisting of two identical HD camcorders.

In order to secure temporal synchronization of the two cameras, a single remote control was employed to activate both of them at the same time instance. The temporal synchronization of the video sequences is further confirmed in the post-processing phase. We checked the left and the right views of each captured video sequence on a frame-by-frame basis and ensured that the two views are temporally aligned. Even though the cameras are carefully lined up, the recorded videos may require rectification to compensate for vertical, horizontal and rotational misalignments that can hamper the perceptual quality of 3D content.

We set the exposure adjustment of the cameras on the manual mode and capture each scene using several different exposure levels. When recording each exposure level, to guarantee that brightness is the only varying parameter, we ensure that the scene relatively stays unchanged in terms of content, object motion and camera movement. This guarantees the consistency needed in our comparisons. The exposure levels are chosen between sufficiently bright to sufficiently dark levels; the number of steps varies depending on the brightness of the scene that is captured. Note that since our cameras do not provide a numerical value for the exposure setting, we divided the exposure range (under expose to over expose) to 6 equal steps for the indoor scenes and 8 equal steps for the outdoor scenes by counting the number of steps on the camera. As expected, outdoor scenes captured during the daytime allow for a wider range of exposures.

B. Subjective Evaluation

The subjective evaluation was performed to investigate the effect of picture brightness on 3D visual perception and then compare it to that of 2D videos. For this experiment, six stereoscopic test sequences were used (two outdoor and four indoor). Fig. 2 shows snapshots of our test sequences. Indoor scenes were captured at six different levels of camera exposures (very dark to very bright), and Fig. 3 shows an example of a scene captured with different brightness levels. Since outdoor scenes allow a wider range of brightness, the number of exposure levels for the outdoor sequences was increased to eight. In both cases, outdoor and indoor, each sequence is approximately 10 seconds long.

The viewing conditions were set according to the ITU-R Recommendation BT.500-11 [8]. Eighteen observers participated in our subjective tests: six females and twelve males, ranging from 23 to 63 years old. All subjects have none to marginal 3D image and video viewing experience. A 65” Full HD 3D display (©Panasonic, Plasma. TC-P65VT25) was used in our experiment. The TV settings were as follows: brightness: 80, contrast: 80, color: 50, R: 70, G: 45, B: 30.

At the beginning of the experiment a training section was provided: the “Running” test sequence was played starting from a very bright exposure to a very dark exposure to help viewers become familiar with the test process and show them the expected range quality-change. Note that “Running” was excluded from the subsequent testing process.

Our tests included two steps: rating the perceptual quality of 3D content in different exposure levels and grading the

quality of the same content in 2D format. In order to maintain similar brightness conditions for both cases (3D and 2D), the left view of the captured video sequences was chosen and displayed to both eyes in the 2D case and the viewers were asked to use the glasses while watching the 2D video sequences. In our tests, subjects were aware if they were watching 2D or 3D videos. Both 2D and 3D subjective tests were conducted at the same time and the same room.

The single stimulus continuous quality evaluation (SSCQE) method is used in our subjective tests [8]. After the training sequence, the viewers were first shown each stereoscopic test sequence in random order of exposure levels. Between two consecutive videos with different exposures, a three-second gray interval is provided for allowing the viewers to rate the perceptual quality of the content and relax their eyes before watching the next video. The perceptual quality reflects whether the displayed scene looks pleasant in general. In particular, subjects were asked to rate a combination of "naturalness", "depth impression" and "comfort" as suggested by Hyunh-Thu et al. [9]. Fig. 4 demonstrates the testing procedure. The scoring bar ranges from 0 – 10 in a continuous scale and a higher score correspond to better quality. Once all the stereoscopic sequences were played, there was a break interval of 5 seconds followed by the 2D sequences. The same process was followed in this case - the same sequences in 2D format

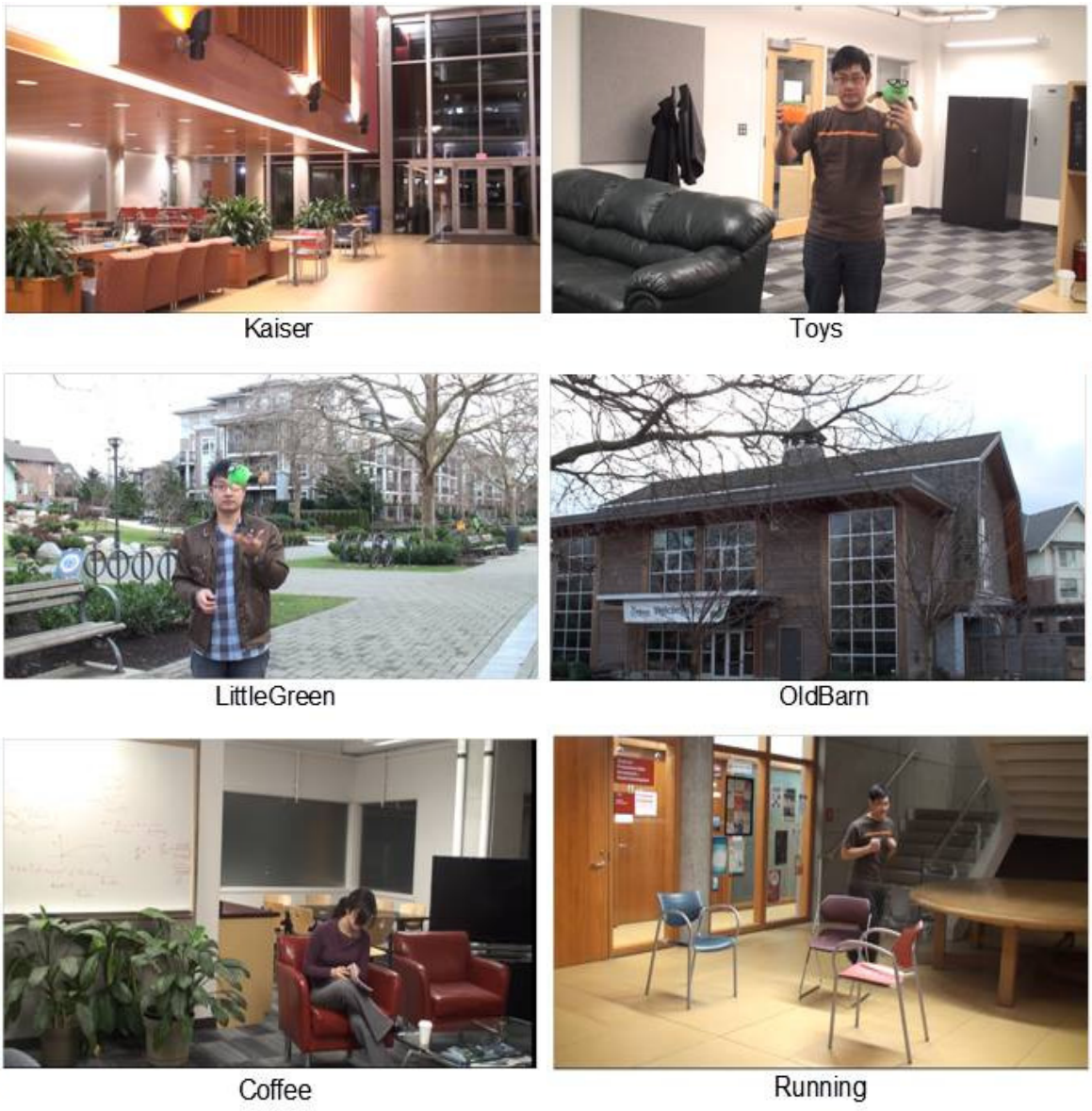


Figure 2. Snapshots of captured video scenes for the first experiment.

were shown in random order of brightness, with 3-second intervals between each sequence - with the viewers asked to rate the video quality of the 2D content.

C. Results and Analysis

After collecting the experimental results, we removed the outliers based on the TU-R Recommendation BT.500-11 [8] (there was one outlier) and then the mean opinion scores from viewers are calculated. Fig. 5 shows the results with 95% confidence interval. A general observation that applies to outdoor and indoor scenes is that the video sequences with brightness levels around the mid-band are more appealing to the viewers in both the 2D and 3D cases (rating scores are more than 6). Note that the exposure range is wider for the outdoor scenes and the highest brightness level for outdoors cannot be reached for the indoor scenes. The reason is that high levels of brightness in indoor scenes will introduce unacceptable degrees of noise caused by the sensitivity of the camera sensors. Regarding the indoor scenes, we observe that for scores above 6 (over which the quality may be regarded acceptable), the quality of the 3D experience increases as the brightness increases. This trend is also observed in the 2D case, with scores following the same pattern. If we would like to conclude something more from these findings, we could say that the 3D experience adds a bit more quality overall, although the difference is not very large.

For the outdoor sequences, we observe that the extreme levels of brightness, very dark and very bright, result in unacceptable quality. For the mid-band section, the quality stays at acceptable levels, with very small variations. The 2D performance is almost identical, with 2D ratings slightly

Figure 3. Different (global) brightness levels in the first experiment.

higher than those of 3D in this case. The main reason for this deviation from the indoor case is the actual content of the outdoor scenes. The objects with depth in the latter case are relatively darker than the background. In fact, a large portion of the background – more so in the "OldBarn" sequence than the "LittleGreen" – is the sky, which in both cases is very bright and lacks detail. According to the viewers, the lack of detailed information in the background (i.e., overexposed sky) drastically reduced the level of depth illusion, hampering the 3D perception when watching stereoscopic content.

III. Experiment Two – Local Change Of Brightness

In this section, we also study the effect of local brightness-change on the perceived 2D and 3D quality. The difference from the previous section is that, for each scene, the brightness of the background is kept relatively consistent while the brightness of the object(s) of interest is changed.

A. *Capturing Setup*

In our study, we intend to investigate the effect of the object's brightness-change with respect to the background on the perceived visual quality in both 2D and 3D cases. For this comparison, 3D videos of indoor and outdoor scenes are captured using stereo cameras. Obtaining 2D videos is trivial since they can be derived directly from the right or left view of the 3D counterparts. Below we present the setup of our 3D capturing.

For each scene, the brightness of the object(s) changes from an under-exposed to an over-exposed level, while the brightness of the background is adjusted to a normally exposed level (not over/under exposed) and is kept relatively unchanged for all the different recordings of the same scene.

Figure 4. Our subjective test procedure.

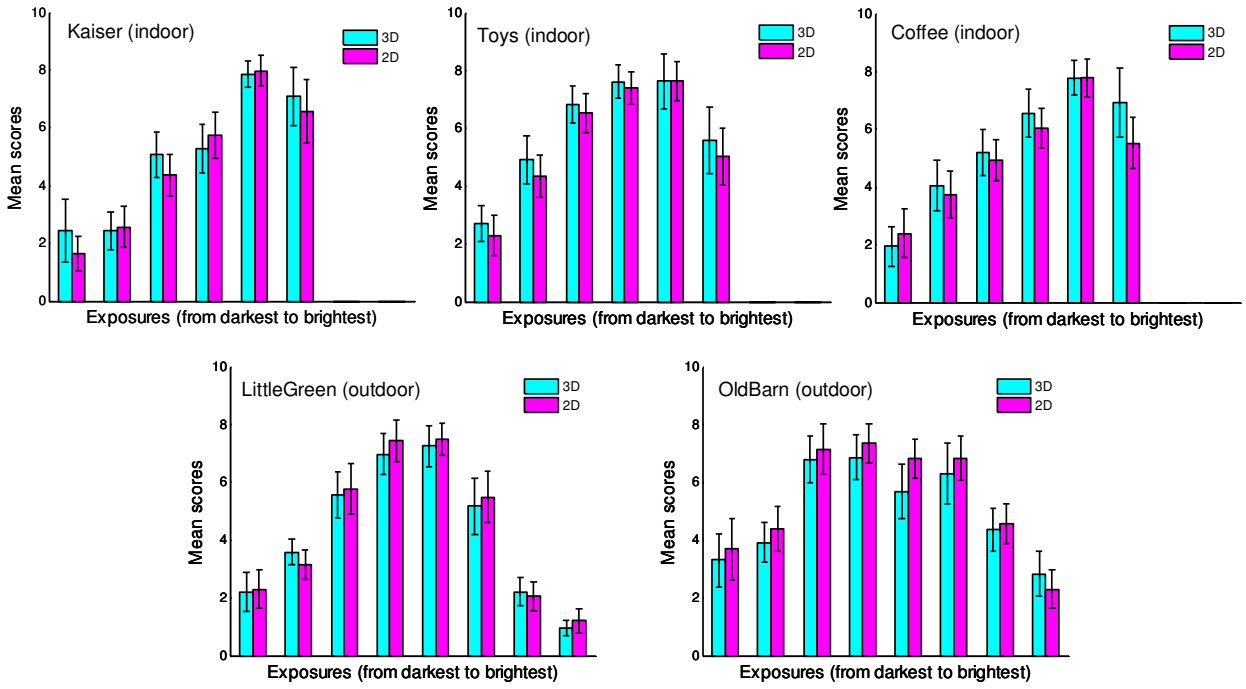


Figure 5. Subjective test results of the first experiment.

The same capturing set up as the one described in Section II.A for the first experiment (Fig. 1) was used. For the indoor scenes, the brightness of the object(s) was changed by using a dimmable 1000W fluorescent video light source (FloLight FL-220AW). For the outdoor scenes, since the emitted light from the light source was insufficient for changing the brightness of the object(s) (due to the presence of sunlight), we used a collapsible circular reflector disc with multiple impacts to reflect different levels of sunlight on the object(s). Fig. 6 shows our capturing setup, the light source and the reflector used in our experiments. In general, capturing outdoor scenes was much more challenging compared to indoor scenes, due to the presence of sunlight and the change of weather conditions, which kept altering the background brightness.

For our study, at each brightness level, the luminance of the object(s) and background is measured using a multifunction light meter (Sekonic L-758Cine). We measure luminance since it indicates how much luminous power is perceived by the human eye when viewing the surface from a particular angle. To measure the brightness of object(s), we took the measurement at each brightness level using a specific spot on the object(s) which subjective tests shown to be the focus of attention. For example, in the case of the object of interest being a human, all measurements were taken from face. Regarding the background, we conducted several measurements, we made sure that the fluctuations were very small, and then took the average value.

B. Subjective Evaluation

For this experiment, six stereoscopic test sequences (two outdoor and four indoor) and one demo video were captured using the stereo camera setup described previously. Fig. 7 shows a snapshot of our test sequences. For each scene the camera exposure is adjusted such that the background area is neither overexposed nor underexposed. The brightness of the object(s) between consecutive contrast levels are different by about 2/3 stop, which was achieved by measuring the luminance of the object(s) using a luminance meter. Then, the brightness of the object(s) is changed from an underexposed level to an overexposed level within multiple steps, with the brightness-change remaining visually differentiable (see Fig. 8). In both cases, outdoor and indoor, each sequence is approximately 10 seconds long. For each scene recording, we ensure that while the object's brightness changes, the content of the scene and the background luminance remain unchanged.

To quantify the perceived quality of the 3D experience at different levels of brightness, we performed subjective quality assessment tests. The viewing conditions of our subjective test were set according to the ITU-R Recommendation BT.500-11 [8]. Eighteen observers participated in our subjective tests: seven females and eleven males, ranging from 23 to 60 years old. All subjects had none to marginal 3D image and video viewing experience. The same display as of our previous study was used for this experiment: a 65" Full HD 3D display (©Panasonic, Plasma, TC-P65VT25) and the subjective-test procedure was similar to our first experiment (see Section II, part B).

C. Results and Analysis

After collecting the experimental results, we checked for the outliers based on the TU-R Recommendation BT.500-11 [8] and then the mean opinion scores (MOS) from viewers were calculated. Fig. 9 shows the average perceptual 3D quality (MOS) versus brightness of the object(s) for all six stereo sequences. As it can be observed, the acceptable brightness level for objects in outdoor scenes is much higher than those in indoor scenes, due to the presence of sunlight. Here, the numerical value of object(s) brightness could not be used as a guideline for capturing high quality 3D content, in other words, we can not conclude that there is a certain range of brightness of the object(s) that results in acceptable 3D quality (where MOS is greater than 6). The reason is that the measured brightness is not a good indicator of what the camera has captured. In order to remove the camera effect, we need to avoid using the brightness and instead calculate the contrast between the objects of interest and the background. There are many ways of measuring contrast. In this study we chose to use the Weber approach [10], which is one of the most commonly used in the field:

$$\text{Contrast} = \frac{L_o - L_b}{L_b} \quad (1)$$

Figure 6. Capturing setup for the second experiment.

where L_o is the luminance of an object and L_b is the average luminance of the background.

Please recall that in this experiment, the brightness of background is kept relatively constant and only the brightness of object(s) of interest is changed for each scene. This allows us to create videos with the same content but different contrast between the object(s) of interest and the background, a necessary feature for our subjective tests, as we do not want the content itself to affect the viewers' decision.

Fig. 10 shows the average subjective scores for quality of 3D content versus contrast for all six sequences. A general observation that applies to both outdoor and indoor scenes is that the stereo video sequences with Weber contrast levels of *-0.35* to *0.55* between the object and background are more appealing to the viewers (these correspond to rating scores above 6, which may be regarded as acceptable quality). Note that although the visually acceptable range of object's brightness (MOS over 6) is higher for the outdoor scenes compared to that of the indoor scenes (as shown in Fig. 9) the range of contrast that ensures high 3D quality is similar for both cases. It is also observed that low scores are associated with high contrast scenes, which in Fig. 10 appear at both ends of the horizontal axis, as contrast here is the difference between the objects' brightness and that of the background.

Fig. 11 shows the average perceptual quality versus contrast for all six scenes. Each point in the figure denotes the mean opinion score for a specific scene at a specific contrast level. The blue solid line is the average MOS over all the scenes, and the upper and lower red dash lines denote the 95% confidence interval. It can be observed that the

Figure 7. Snap shot of test sequences for the second experiment.

scores obtained from rating the 3D and 2D videos are following the same pattern, especially in the range of acceptable video quality (scores > 6), which indicates that the contrast or the brightness of the object of interest has almost equivalent influence on the perceptual quality for acceptable 3D and 2D videos. We also observe that scores for 2D are higher than those for 3D at the left end of the curve (where the object is much darker than the background). This is because crosstalk (ghosting) artifacts are severe whenever the background brightness is much higher than that of the objects of interest.

To further verify that the effect of contrast on the quality of 3D and 2D videos is similar, we performed the statistical significance test. Hogg's one-way analysis of variance (ANOVA) [11] is applied to 2D and 3D scores for each scene. We state the null hypothesis as: there is no significant distinction between the collected values for the perceptual quality of 2D and that of 3D. The higher the returned p value is (which considers the difference in average and variance as well as the sample size), the less significant is the difference. A typical threshold of the p value that rejects the null hypothesis is 0.01 or 0.05. The p values for six scenes are 0.3990 (Reading), 0.7663 (Interview), 0.9068 (Clapping), 0.4104 (SlowMotion), 0.8228 (McMillan) and 0.8499 (Painting) respectively, which are far above the threshold. Therefore, based on our experiment at each contrast level, there is no statistically significant difference between the perceptual quality of 3D and 2D videos of similar scenes. That is, given a large sample size, 3D and 2D videos should have virtually identical preferred contrast levels.

Based on our observations, 3D content captured according to the recommended contrast range can provide high quality 2D experience. In this case, if the system involves the use of glasses, the display should adjust the

Figure 8. Same scene with different (local) brightness levels of the object.

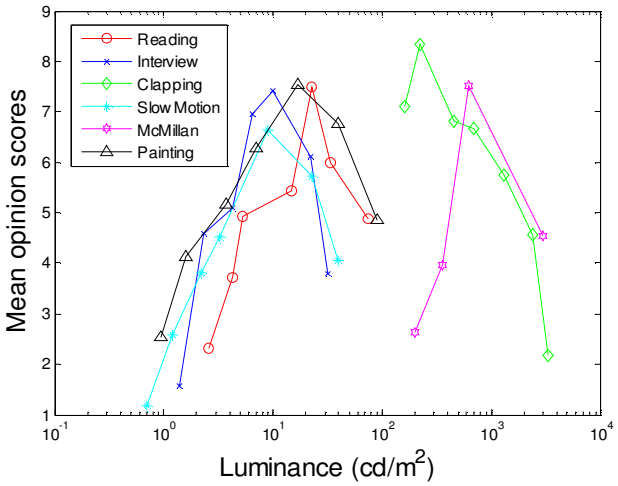


Figure 9. Perceptual 3D quality score versus measured brightness.

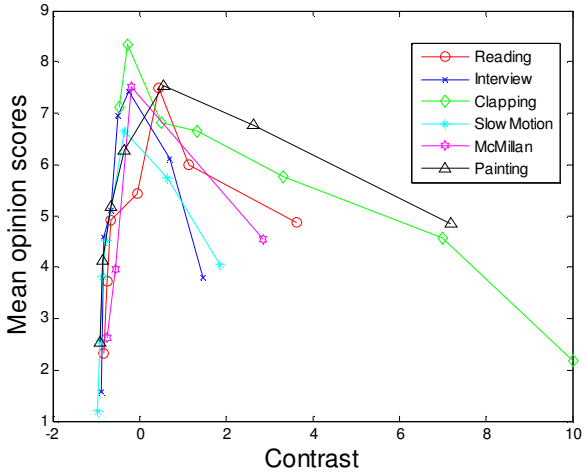


Figure 10. Perceptual 3D quality score versus the contrast between object and background.

contrast of the 2D stream, to compensate for the diming effect of the 3D glasses.

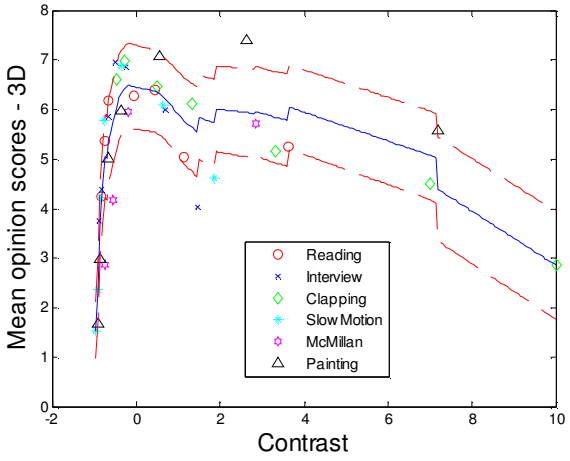
IV. DISCUSSION

Our psychophysical experiment involved considerable human power, including the subjects that took the tests and the researchers who prepared and organized the tests. These efforts could be reduced if a reliable 3D quality metric is available. Presently, the community is putting a considerable effort towards achieving this goal, but we are still far from an acceptable solution.

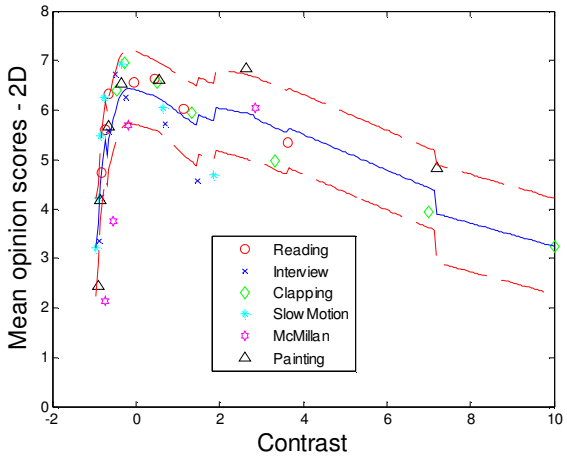
Our finding that brightness has similar effects on 3D and 2D videos is different from the practical observation in [5]. The latter suggests capturing 3D videos with higher light settings than 2D content in order to achieve good perceptual quality. This contradiction stems from the fact that we use 3D glasses in both 2D and 3D subjective tests so that the light reduction for 3D and 2D is similar, whereas in [5] the light reduction was considered only for 3D.

Our results on the relationship between 2D and 3D viewing (for both brightness and contrast) can be generalized for existing 3D displays, although the display systems deliver different amount of light when they display 3D content compared to 2D. In our study, this is due to the fact that both the 2D and 3D viewing tests were conducted while the display was set to the 3D mode and subjects wore 3D glasses during the entire experiment. On the other hand, the preferred range of contrast may vary depending on the displayed luminance range that the 3D viewing system supports. The variation, however, would not be large since contrast is defined as a ratio that is normalized by the background luminance.

The low scores at both ends on the x-axis in Fig. 10 and Fig. 11 are due to the effect of crosstalk. Crosstalk is an



(a) Scores for 3D video



(b) Scores for 2D video

Figure 11. Average perceptual quality versus contrast. Each point denotes the mean opinion score for a specific scene at a specific contrast level. The blue solid line cross the plot is the average MOS over all the scenes, and the upper and lower red dash lines denote the 95% confidence interval.

artifact where light that should be delivered to one eye is "leaked" to the other eye (it is also referred as the "ghosting" artifact) [12]. It is well known that crosstalk artifacts in 3D displays become severe when the contrast is high [13]. Note that all the existing 3D displays suffer from crosstalk. To reduce the crosstalk effect, we picked a 3D TV system, which based on subjective tests offers the best crosstalk reduction performance.

Eye strain and headaches have been noted as some of the most important factors that affect the overall 3D quality [14]. In our tests, subjects were not explicitly asked to evaluate this factor. Properly identifying eye strain and headache requires a long-time viewing of a stimulus [15]. This is not practical in our current design, since it would make the testing session too long, and as a result it may affect the reliability of the collected data.

V. Conclusion and Future Work

The era of user-centric multimedia has already begun, and quality plays a central role in it. Attention to the quality in 3D content case is even more important since low-quality 3D videos can produce eyestrain, headache, and generally unpleasant viewing experience for the viewers. Brightness is one of the important factors that affect the visual comfort and quality of 3D videos. In this work, we investigated the effect that global and local brightness-changes have on quality of 3D experience and its relation to the 2D scenario. This was done by performing extensive subjective quality assessments to quantify the perceived quality of the 3D and 2D experience at different levels of brightness.

Our work contained two studies: 1) global change of brightness, which is achieved by setting different exposure times on the camera, and 2) local change of brightness, where the brightness of the background was kept relatively constant and the object's brightness was changing using an external light source and/or a reflector disk. Subjective evaluations were performed, with the subjects being asked to rate the 3D and 2D perceptual quality of each sequence.

A conclusion from both studies is that marginal difference is found in the brightness effects between 3D and 2D videos. According to our results from the second study, a general observation that applies to outdoor and indoor scenes is that the stereo video sequences with Weber contrast levels of *-0.35* to *0.55* between the object(s) and the background are more appealing to the viewers. In summary, content producers may improve the overall 3D quality of experience by adjusting the brightness of the objects of interest in a scene to ensure that the Weber contrast the objects and background falls within the suggested range levels. Film industry could use our findings to produce 3D videos with different brightness levels or light-settings.

Our future work include: employ special lighting equipment in order to explore any possible case of brightness for foreground and background in outdoor and indoor conditions, conduct subjective tests with a much larger variety of content, and investigate the effect of different parameters on 3D perception quantitatively. This is towards our ultimate goal, which is providing guidelines for capturing and displaying 3D content as well as developing a 3D quality metric.

Acknowledgment

The authors thank all the subjects that participated in our test.

References

[1] M. T. Pourazad, Z. Mai, and P. Nasiopoulos, "Effect of contrast on the quality of 3D visual perception," Proc. Third International Conference on Creative Content Technologies, CONTENT'11, September 25-30, 2011, Rome, pp. 43-47, IARIA XPS Press, ISBN: 978-1-61208-157-1

[2] G. Jones, D. Lee, N. Holliman, and D. Ezra, "Controlling Perceived Depth in Stereoscopic Images," Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems, vol. 4297, pp. 42–53, 2001.

[3] L. Goldmann, F. D. Simone, and T. Ebrahimi, "Impact of acquisition distortion on the quality of stereoscopic images," in International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, USA, 2010.

[4] D. Xu, L. Coria, and P. Nasiopoulos, "Guidelines for capturing high quality stereoscopic content based on a systematic subjective evaluation," Proc. of IEEE International Conference on Electronics, Circuits, and Systems, ICECS 2010, pp. 166-169, Athens, Greece, 2010.

[5] B. Mendiburu, "3D Movie Making – Stereoscopic Digital Cinema from Script to Screen," Elsevier, 2008.

[6] L.K. Cormack, S.B. Stevenson, and C.M. Schor, "Interocular correlation, luminance contrast and cyclopean processing," Vision Research, vol. 31, no.12, pp. 2195–2207, 1991.

[7] L. M. J. Meesters, W. A. IJsselsteijn, and P. J. H. Seuntiënsm, "A survey of perceptual evaluations and requirements of three-dimensional TV," IEEE Trans. Circuits & System for video technology, vol. 14, no. 3, pp. 381-391, 2004.

[8] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R, Tech. Rep. BT.500-11, 2002.

[9] Q. Hyunh-Thu, P. L. Callet, and M. Barkowsky, "Video quality assessment: from 2D to 3D challenges and future trends," Proc. of 2010 IEEE 17th International Conference on Image Processing, (ICIP), pp.4025-4028, 2010.

[10] R. Shapley and C. Enroth-Cugell, "Visual adaptation and retinal gain controls," Progress in Retinal Research, vol. 3, pp. 263-346, 1984.

[11] R. V. Hogg and J. Ledolter, "Engineering Statistics," New York: MacMillan, 1987.

[12] A. J. Woods, "Understanding Crosstalk in Stereoscopic Displays" (Keynote Presentation) at Three-Dimensional Systems and Applications (3DSA) conference, Tokyo, May 2010.

[13] S. Shestak, D. Kim, and S. Hwang, "Measuring of Gray-to-Gray Crosstalk in a LCD Based Time-Sequential Stereoscopic Display," Proc. of the Society for Information Display (SID 2010), Seattle, pp. 132-135, May 2010.

[14] M.T.M Lambooij, W.A. IJsselsteijn, and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XIV, vol. 6490, Jan. 2007.

[15] M. Pölönen, T. Jarvenpaa, and J. Hakksinen, "Comparison of near-to-eye displays: subjective experience and comfort," J. Display Technol., vol. 6, no. 1, pp. 27–35, Jan. 2010.

www.iariajournals.org

International Journal On Advances in Intelligent Systems

ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

issn: 1942-2644

International Journal On Advances in Security

ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

issn: 1942-2636

International Journal On Advances in Software

ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

issn: 1942-2601