International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications

July 7 - 9, 2020, Graz, Austria

Proceedings

IEEE Austria Section

TU Graz

Content

Foreword / Chairmen's Message

On behalf of the Local Organizing Committee, the International Technical Committee and all the sponsors and patrons, we are delighted to welcome you all to the 3rd International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom) – hosted by the Faculty of Electrical Engineering at Graz University of Technology, Graz, Austria, 7–9 July 2020. Compared to previous events, i.e., CSNDSP 2008, ConTEL 2011, NOC 2013, ICTON 2014, ConTEL 2015, CoBCom 2016, CoBCom 2018 and ConTEL 2019, which took place in Graz so far (since more than 10 years now), we have a new situation in 2020 regarding the COVID-19 breakout. The COVID-19 has changed our lives and we have learned, more or less, to cope with this sort of virus, because it will not disappear in the next future. We had to find new methods of presentations and to investigate countermeasures to keep small conferences alive. Such small conferences are particularly important for young scientists and PhD students.

For this reason we confirmed to hold this conference as a virtual one, which was allowed by IEEE in the case of the COVID-19 pandemic. In line with the earlier events in Graz, we will host virtual sessions on the topics of Networks and Protocols, Future and Innovative Communication Technologies, Satellite and Space Communications, Photonic and Optical (Wireless) Communications Systems and Networks, Multimedia Applications as well as RF Engineering and Antennas, organized by the experts in the field to whom we are very grateful. The program also includes internationally well-known keynote speakers presenting overviews of current research in future topics and innovative areas.

The 3rd International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom), hosted by Graz University of Technology, Austria, was in 2016 a newly generated international conference and repeated in 2018. The CoBCom is a small conference, but we are really proud that we have so much international participants and that we received the technical co-sponsorship of the IEEE / IEEE Austria Section also for 2020. On behalf of the technical and organizing committees, we are presenting the CoBCom 2020 program, providing a platform for scientists, industry, operators, and service providers to share ideas, discuss recent advances, exchange their R&D experience, and present state-of-the-art techniques and future technologies. The CoBCom 2020 technical program comprises keynote talks, a general track, sessions related to newest technologies, smart cities and multimedia services, as well as an „International CEEPUS and ERASMUS Workshop" on microwave technologies, radars, remote sensing and communications with three sessions (General CEEPUS Special Session, Special Session on Remote Sensing and Radars, Special Session on Microwave Technologies and Communications) strongly connected to regional companies and neighbour country's institutions.

At CoBCom 2020, keynote talks are presented in form of plenary sessions. The first keynote speaker is Dr. Matevž Pogačnik (University of Ljubljana). His speech about "Trends and Challenges in Virtual Reality Services for 360° Video Delivery and Consumption" addresses the technology and service trends in the virtual reality domain through a review of technical properties of 360° cameras and virtual reality head-mounted-displays, new content formats

and upcoming approaches to high-bandwidth 360° video streaming. The second keynote about “5G and Cyber Security – how to make 5G networks more secure” is presented by Dr. Kurt Reichinger form the Technical Division at the Austrian Regulatory Authority for Broadcasting and Telecommunications. The third keynote, entitled “Environmental Sustainability of Information and Communication Technology (ICT) for Smart Grids”, is given by Prof. Slavisa Aleksic, Hochschule für Telekommunikation, Leipzig, Germany. In his talk, he discusses the environmental sustainability of the ICT equipment for smart grids that is involved in the advanced metering infrastructure and home area network applications.

Due to the COVID-19 crisis, the Industrial Workshop (like at CoBCom 2016 and 2018) could not be organized. This loss of the Industrial Workshop in 2020 generated also a big lack of patrons, because industry likes to sponsor events with possible future customers. So we are satisfied that we could collect 41 paper entries via our EDAS system. After the review process, we selected 29 papers for presentation at CoBCom 2020 (among them a few with major revision). At the submission deadline for the final manuscript, we accepted 28 contributions. These papers were registered and are ready to be presented. For papers submission and reviewing process we used the EDAS system. Each paper was evaluated by at least three independent reviewers with respect to their technical content, novelty, originality, and presentation. All the 28 accepted and registered contributions will be published via IEEE Xplore in case they are presented at the conference. In contrast to earlier conferences in Graz we had less patrons as a consequence of the COVID-19 situation.

Our sincere thanks go to the people whose cooperation and help made this conference possible, in particular to the members of the international Organizing and Technical Program Committees, but also to people responsible for local support, to all the authors submitting their work to CoBCom 2020 and, last but not least, the reviewers for their time and effort spent in this context. The technical co-sponsorship of the IEEE Austria Chapter is very much appreciated and hereby acknowledged. For participants (specially the CEEPUS workshop organisers Peter Planinšič and Galia Marinova), staying in Graz within the time of the virtual conference (caused by the COVID-19), we have also organised two social events (Guided City Tour with Lord Mayor Reception and Invitation to Heurigen / Buschenschank by the Governor of Styria). In this case special thanks go to the support and warm hospitality extended to the conference participants by the Province of Styria and the City of Graz (including Graz Tourism).

Finally, we wish all the participants a rewarding meeting, exciting presentations, fruitful discussions, a nice working atmosphere, and that the few on-site participants also enjoy the social program!

Erich Leitgeb, Wilfried Gappmair, Maja Matijašević - General Chairs

Franz Teschl, Andrej Kos, Dušan Gleich, Mariusz Głąbowski - TPC Chairs

Committees

General Chairs
Erich Leitgeb, *Graz University of Technology, Austria*
Wilfried Gappmair, *Graz University of Technology, Austria*
Maja Matijašević, *University of Zagreb, Croatia*

Program Chairs
Franz Teschl, *Graz University of Technology, Austria*
Andrej Kos, *University of Ljubljana, Slovenia*
Dušan Gleich, *University of Maribor, Slovenia*
Mariusz Głąbowski, *Poznan University of Technology, Poland*

IEEE Liaison
Wolfgang Bösch, *Graz University of Technology, Austria*

Organizing Committee Chairs
Reinhard Teschl, *Graz University of Technology, Austria*
Alice Reinbacher-Köstinger, *Graz University of Technology, Austria*

Publication Chairs
Franz Teschl, *Graz University of Technology, Austria*
Thomas Bauernfeind, *Graz University of Technology, Austria*

Publicity/Web Chair
Thomas Plank, *Graz University of Technology, Austria*

International Networking & Education Chairs (CEEPUS, ERASMUS, COST)
Peter Planinšič, *University of Maribor, Slovenia*
Galia Marinova, *Technical University of Sofia, Bulgaria*
Vera Markovic, *University of Nis, Serbia*
Pirmin Pezzei, *Graz University of Technology, Austria*

Technical Program Committee
Slaviša Aleksić, *Vienna University of Technology, Austria*
Hovik Baghdasaryan, *National Engineering University of Armenia, Armenia*
Dinko Begušić, *University of Split, Croatia*
Michael Bergmann, *Eutelsat, Paris, France*
Janez Bester, *University of Ljubljana, Slovenia*
Horst Bischof, *Graz University of Technology, Austria*
Janos Zoltan Bito, *Budapest University of Technology, Hungary*
Wolfgang Bösch, *Graz University of Technology, Austria*
Rasa Bruzgiene, *Kaunas University of Technology, Lithuania*
Carlo Capsoni, *Politecnico di Milano, Italy*
Luis M. Correia, *IST/INOV-INESC – Univ. of Lisbon, Portugal*
Bernd Eichberger, *Graz University of Technology, Austria*
Uwe Fiebig, *German Aerospace Center (DLR), Germany*
Michael Gadringer, *Graz University of Technology, Austria*

Michael Gebhart, *EPCOS Ohg, Austria*
Zabih Ghassemlooy, *University of Northumbria, Newcastle, UK*
Mariusz Głąbowski, *Poznan University of Technology, Poland*
Jasmin Grosinger, *Graz University of Technology, Austria*
Steve Hranilovic, *McMaster University, Canada*
Irena Jurdana, *University of Rijeka, Croatia*
Gorazd Kandus, *Jožef Stefan Institute, Slovenia*
Ali M. Khalighi, *Institut Fresnel, Marseille, France*
Andrej Kos, *University of Ljubljana, Slovenia*
Dragana Krsitc, *University of Nis, Serbia*
Michael Logothetis, *University of Patras, Greece*
Ignac Lovrek, *University of Zagreb, Croatia*
Dražen Lučić, *Croatian Chamber of Commerce, HGK, Croatia*
Marian Marciniak, *National Institute of Telecommunications, Poland*
Antonio Martellucci, *ESA/ESTEC,The Netherlands*
Maja Matijašević, *University of Zagreb, Croatia*
Vjekoslav Matić, *Infineon, Graz, Austria*
Branko Mikac, *University of Zagreb, Croatia*
Wai Pang Ng, *University of Northumbria, Newcastle, UK*
Peter Pocta, *Zilina University, Slovakia*
Zbynek Raida, *Brno University of Technology, Czech Rep.*
Vladimir Rastorguev, *MAI, National Research University, Russia*
Peter Rössler, *Technikum Wien, Austria*
Filiz Sari, *Aksaray University, Turkey*
Nikolaus Schmitt, *Airbus (EADS), Ottobrunn, Germany*
Sajid Shaikh Muhammad, *National University of Computer & Emerging Sciences (FAST – NUCES), Pakistan*
Andreas Springer, *Johannes Kepler University Linz, Austria*
Victor Sucic, *University of Rijeka, Croatia*
Franz Teschl, *Graz University of Technology, Austria*
Jan Vitasek, *Technical University of Ostrava, Czech Rep.*
Otakar Wilfert, *University of Technology Brno, Czech Rep.*
Peter Winzer, *Bell Labs, USA*
Harald Witschnig, *Infineon Graz, Austria*
Horst Zimmermann, *Vienna University of Technology, Austria*

Reviewers

Alen Hrga
Andreas Strasser
Andrej Kos
Athanasios Goulas
Bernd Eichberger
Bernhard Schrenk
Christian Elgaard
Daniel Kraus
David Veit
Devi Vinayak Siva Rama Krishna Koilada
Dinko Begusic
Dušan Gleich
Filiz Sari
Funmilayo Offiong
Galia Marinova
Harald Witschnig
Helmut Schreiber
Horst Zimmermann
Irena Jurdana
Ivan Russo
Jan Köhler
Jan Latal
János Bitó
Jugoslav Joković
Klaus Kainrath
Konrad Diwold
Lucie Hudcova
Luis Correia
Madhukar Chandra
Maja Matijasevic
Marian Marciniak
Mario Kusek
Mariusz Glabowski
Matevž Pogačnik
Michael Gebhart
Michael Logothetis
Mohammad-Ali Khalighi
Nikolaus Schmitt
Peter Barcik
Peter Mandl
Peter Planinsic
Peter Pocta
Peter Rössler
Peter Winzer
Philipp Ortner
Pirmin Pezzei
Rasa Bruzgiene
Reinhard Teschl
Reinhard Zeif
Slavisa Aleksic
Stephan Bernhart
Tatjana Nikolic
Thomas Plank
Uwe-Carsten Fiebig
Victor Sucic
Wang Lianmei
Wasiu Popoola
Zbynek Raida
Ziad Hatab

Programme Overview

Day / Time	Tuesday, July 07	Wednesday, July 08	Thursday, July 09
09:00	**Conference Opening**		
09:15			
09:30	***Keynote Talk I Matevž Pogačnik***	***Keynote Talk III Slaviša Aleksić***	**CEEPUS Workshop**
09:45			
10:00			
10:15	**Satellite and Space Communications**	**General Topics I**	
10:30	(3 papers)	(3 papers)	
10:45			
11:00			
11:15			
11:30	**Photonics and Optical Communications**	**General Topics II**	
11:45			
12:00			
12:15	(4 papers)	(6 papers)	
12:30			
12:45			
13:00			
13:15			
13:30			
13:45			
14:00	***Keynote Talk II Kurt Reichinger***	**CEEPUS Workshop**	
14:15			
14:30	**Antennas and Microwave Engineering**		
14:45			
15:00	(3 papers)	(5 papers)	
15:15			
15:30			
15:45			
16:00	**IoT, Networks and Protocols**		
16:15			
16:30			
16:45	(4 papers)		
17:00		**Buschenschank Sattler**	
17:15			
17:30			
17:45			
18:00			
18:30	**Guided City Tour**		
18:45			
19:00			
19:15			
19:30			
19:45			
20:00	**Welcome Reception Kunsthaus Graz**		

Sessions

Tuesday, 07th of July

9:00 – 11:00

Conference Opening

Keynote Talk I **Trends and challenges in virtual reality services for 360° video delivery and consumption** – Matevž Pogacnik

Satellite and Space Communications Chair: Wilfried Gappmair	**Page**
Thermal Vacuum Tests and Thermal Properties on ESA's OPS-SAT Mission Authors: Manuel Kubicka; Otto Koudelka; David Evans; Reinhard Zeif; Max Henkel; Andreas Hörmer	**17**
From OPS-SAT to PRETTY Mission: A Second Generation Software Defined Radio Transceiver for Passive Reflectometry Authors: Reinhard Zeif; Andreas Hörmer; Manuel Kubicka; Max Henkel; Otto Koudelka	**24**
A GPS Patch Antenna Array for the ESA PRETTY Nanosatellite Mission Authors: Reinhard Zeif; Andreas Hörmer; Manuel Kubicka; Max Henkel; Otto Koudelka	**32**

11:30 – 13:00

Photonics and Optical Communications Chair: Erich Leitgeb	**Page**
DF Relayed OOK and PAM FSO Links with Turbulence and Time Jitter Authors: Panagiotis Gripeos; Hector Nistazakis; George Roumelas; Vasilis Christofilakis; Andreas Tsigopoulos; George S Tombras	**39**
Implementation of Intelligent Modulator into the Luminaire of Public Lighting Based on the OOK Modulation with Bias-Tee Authors: Stanislav Hejduk; Tomas Stratil; Jan Latal; Ales Vanderka; Lukas Hajek; Jakub Kolar	**46**
3D Visible Light Positioning of an Angle Diverse Receiver Based on Track Analysis Authors: Andreas Weiss; Franz Wenzl; Claude Leiner; Felix Lichtenegger; Christian Sommer	**51**
Modelling the Refractive Index Structure Parameter: A ResNet Approach Authors: Christopher Lamprecht; Pasha Bekhrad; Hristo Ivanov; Erich Leitgeb	**58**

Tuesday, 07th of July

14:00 – 15:30

Keynote Talk II **5G and Cyber Security – how to make 5G networks more secure** – Kurt Reichinger

16:00 – 17:30

Wednesday, 08th of July

09:30 – 11:00

Keynote Talk III **Environmental Sustainability of Information and Communication Technology (ICT) for Smart Grids** – Slaviša Aleksic

11:30 – 13:30

Wednesday, 08th of July

14:00 – 16:00

CEEPUS Workshop Chairs: Erich Leitgeb / Galia Marinova	**Page**
Optimization of Thick BoR Monopole Antennas Using Differential Evolution Authors: Marko Radovic; Gorazd Lešnjak; Peter Kitak; Peter Planinsic	**160**
DVFS Technique on a Zynq SoC-Based System for Low Power Consumption Authors: Marsida Ibro; Galia Marinova	**164**
Design of a Sampling Mixer for Use in UWB Radar Applications Authors: Marko Malajner; Dušan Gleich	**169**
Interference Classification for IEEE 802.15.4 Networks Authors: Uros M Pesovic; Sladjana Djurasevic; Vanja Lukovic; Peter Planinsic	**173**
USRP Implementation of a Ground Penetrating Radar Using a Combination of Stepped Frequency and OFDM Principles Author: Venceslav Kafedziski; Sinisha Pecov; Dimitar Tanevski	**177**

Remark: The presenting authors are underlined in this programme.

Keynote Talks

Prof. Matevž Pogačnik: Trends and challenges in virtual reality services for 360° video delivery and consumption

Abstract:

In recent years we are witnessing significant advances in the domain of virtual (VR) and augmented (AR) reality solutions. The application domains where VR and AR technology can be used vary from industry, entertainment, education, health care, tourism, etc. One of the challenges of the VR industry is to improve the perceptual quality of the video content, as the standard 2D video formats with 4K or 8K resolutions do not satisfy the quality requirements in the case of 360° videos. While the 360° video production equipment as well as the head-mounted displays (HMD) for the presentation of VR content have significantly improved the resolution and quality of 360° videos, the problem of the increasing bandwidth, required for 360° video streaming, remains.

In this talk, we will discuss the technology and service trends in the VR domain through a review of technical properties of 360° cameras and VR head-mounted-displays, new content formats and upcoming approaches to high-bandwidth 360° video streaming.

Biography:

Assoc. Prof. Dr. Matevž Pogačnik holds a Ph.D. in electrical engineering. He heads the multimedia section of LTFE and LMMFE laboratories at Faculty of electrical engineering, University of Ljubljana. His research and scientific work is focused on development of interactive multimedia services for different devices with a special emphasis on UCD of applications including different interaction modalities for application control. One of his main research interests is design and evaluation of user interfaces using AR and VR technologies, focusing on users with disabilities or users in the rehabilitation process. He is an active member of the IEEE organization.

Dr. Kurt Reichinger: 5G and Cyber Security – how to make 5G networks more secure

Abstract:

In March 2019, the European Commission published a recommendation on cybersecurity of 5G networks which was followed by intense work across the European Union. Member States carried out risk analyses, produced a coordinated European risk assessment and formulated risk mitigation measures. Finally, a common European toolbox for 5G cybersecurity was published in January 2020. In Austria, a large amount of measures from the toolbox is now getting implemented by means of a Network Security Ordinance. The talk recapitulates the work done so far and introduces the measures set to increase security of 5G networks.

Biography:

Kurt Reichinger is Head of Technical Division at the Austrian Regulatory Authority for Broadcasting and Telecommunications. Particular areas of interest are Next Generation Fixed and Mobile Networks, Network Security and Quality of Service.

Kurt Reichinger holds a PhD in Computer Science and an MSc in Telecommunications Engineering both from Vienna University of Technology, Austria; in addition, he received a diploma in Business Administration from Hagen University, Germany.

Prof. Slaviša Aleksić: Environmental Sustainability of Information and Communication Technology (ICT) for Smart Grids

Abstract:

Integration of information and communication technologies (ICTs) into systems for electricity generation, distribution and consumption is inevitable for the deployment of smart grids. Consequently, in order to enable potential efficiency gains of smart grids, a large amount of ICT equipment has to be integrated in various domains, especially in the customer and distribution domains. This additional ICT equipment requires electricity for operation and contributes to the overall electricity consumption. Additionally, it has to be manufactured, transported, installed and, later on disposed or recycled, which unavoidably leads to increased environmental sustainability issues.

Despite the undoubted potentials of smart grids for increasing the overall efficiency of electricity grids in several domains, there is a need to estimate environmental impacts of the additional ICT equipment required to make the vision of smart grids a reality.

In this talk, we will discuss environmental sustainability of the ICT equipment for smart grids that is involved in the advanced metering infrastructure (AMI) and home area network (HAN) applications. The environmental sustainability is analyzed by means of the exergy-based life cycle assessment (E-LCA), which is an approach based on the second law of thermodynamics that takes into consideration the entire lifetime of the equipment. We will indicate the components and life-cycle stages with the highest impact on the environmental sustainability and discuss potentials for improvement.

Biography:

Slavisa Aleksic received both Dipl.-Ing. (M.Sc.) and Dr. techn. (Ph.D.) degrees as well as venia docendi teaching authorisation (habilitation) from the Vienna University of Technology, Austria. Currently, he is a professor at the Hochschule für Telekommunikation (HfTL), Leipzig, Germany, where he is responsible for teaching and research within the broad area of telecommunication technologies and networks with emphasis on network design, management, and performance evaluation. Slavisa Aleksic is author or co-author of more than 130 scientific publications including book chapters, papers in peer-reviewed scientific journals, and contributions to internationally recognized conferences. He has been involved in a number of projects related to communication systems and networks including projects funded by the European Union, Austrian Science Fund (FWF), Austrian Research Promotion Agency (FFG), and in collaboration with several companies.

Prof. Aleksic is a Senior Member of the IEEE and a member of the OVE, the MRS and the IEICE. He serves as a reviewer of numerous reputed journals and book series published by e.g. IEEE, OSA, EURASIP, OSA, ELSEVIER, SPRINGER, TAYLOR & FRANCIS and a member of the Technical Programme Committee of a dozen of international conferences (IEEE, IFIP, IEE, OSA, IAIRA). He received several international and national awards and grants such as four best paper awards, the biggest Austrian business plan award i2b (second level), and the grant of the Austrian Federal Ministry of Education, Science and Culture.

Presented Papers

Thermal Vacuum Tests and Thermal Properties on ESA's OPS-SAT mission

Manuel Kubicka
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
manuel.kubicka@tugraz.at

Otto Koudelka
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
koudelka@tugraz.at

David Evans
Advanced Concepts and ManagementSupport Office
European Space Operations Centre
Darmstadt, Germany
David.Evans@esa.int

Reinhard Zeif
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
reinhard.zeif@tugraz.at

Maximilian Henkel
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
henkel@tugraz.at

Andreas J. Hörmer
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
hoermer@tugraz.at

***Abstract*— OPS-SAT is a 3U CubeSat, designed for versatile use as an experimental platform for industry and universities, to demonstrate new operational concepts and prototype software in a real space environment. The satellite offers numerous payloads alongside the satellite bus, all of which might be used by an OPS-SAT experiment. The unpredictable nature of experiments with respect to the use of payload components raises certain unknowns, in particular concerning power consumption. As a result, the thermal behaviour throughout the satellite depends largely on which of the several on-board experiments and the associated payloads are switched on. OPS-SAT offers a variety of communication modules, such as a UHF transceiver, an S-Band transceiver, a Software Defined Radio (SDR), an X-Band transmitter and an optical receiver. The peak power consumption of OPS-SAT may exceed 30 watts during high power experiments. The S-Band transceiver consumes up to 10 watts during ground station passes and the so-called Satellite Experimental Processing Platform (SEPP), the heart of OPS-SAT experiments, consumes up to 8 watts constantly. This work provides an overview of the design and the thermal considerations on OPS-SAT and the results of the thermal vacuum (TVAC) test campaign. The results yield an average thermo-optical emissivity of 0.79 to 0.84 and the thermal power distribution on the spacecraft surface, and demonstrate the special case of the thermally isolated S-Band and X-Band patch antennas. Based on the derived results, predictions can be made about the thermal behaviour during various load cases and during periods with an active S-Band transmitter.**

Keywords—OPS-SAT, CubeSat, Thermal Vacuum Test

I. INTRODUCTION

It is highly desirable for space missions, that the components in use, the underlying software and the operational concepts are reliable, safe and provide flight heritage. Therefore, anything new has to undergo an extensive qualification process before it can be deemed trustworthy for space flight. Without any flight heritage, it is therefore difficult to become part of a mission and to gain flight heritage in the first place. To break this cycle, ESOCs Advanced Operations Concepts Office came up with the idea of the Operation's Satellite "OPS-SAT" [1]. OPS-SAT is a 3U CubeSat, built by Graz University of Technology (TUG) and was launched in December 2019. Designed as an on-orbit testbed, OPS-SAT offers a dedicated experimental platform, completely separated from the main satellite bus. Industry or universities can apply to become an OPS-SAT experimenter and make full use of the spacecraft's numerous payloads, to test prototype software and new operational concepts in a real space environment. At the heart of OPS-SAT experiments lies the so-called Satellite Experimental Platform (SEPP), a dual-core ARM CPU, coupled with an FPGA, providing Linux as a default operating system. Experiments are started on the SEPP and can take control over almost any aspect of the spacecraft. To communicate with the experiments and to bring results back to ground, OPS-SAT offers an S-Band transceiver, an X-Band transmitter, an optical receiver, a Software Defined Radio (SDR) receiver and a retroreflector. A fine Attitude Determination and Control System (ADCS) is available for experiments and an HD camera can be used for all kinds of purposes. In terms of sensors, OPS-SAT provides six photodiodes, a Fine Sun Sensor (FSS), a magnetometer, a gyro, a GPS receiver and a star tracker. If an experiment fails, one of the two redundant On-Board Computers (OBCs) takes over and keeps control of the spacecraft. Combining all of this in a 3U structure, OPS-SAT eventually requires high amounts of power and can generate a lot of heat, depending on the currently active experiment or use case. The spacecraft is almost fully covered with solar arrays, with the exception of the antennas and other peripherals, e.g. sensors and optical lenses. Two double deployable solar arrays on the long side of the structure provide additional power. The peak power production and the peak power consumption lie in the same order of magnitude and can both exceed 30 watts. The sun-synchronous orbit with only a short eclipse season during November to mid February results in substantial amounts of available power but puts some demand on operational strategies to keep the spacecraft thermally safe.

Within the scope of this work, an overview of the mechanical and thermal design of OPS-SAT is provided, and the results of the Thermal Vacuum Tests (TVAC) are shown. The thermal relation of the spacecraft bus and the payloads was determined in the course of the Thermal Vacuum (TVAC) test campaign. The tests were performed under uniform ambient conditions, i.e. no sun simulator or other heat source

was used. Instead, the shroud of the TVAC chamber was heated and cooled, so that the spacecraft components could stabilise at their maximum respective upper and lower temperatures. As TVAC tests are very time consuming, it is not feasible to consider all possible experiment configurations. Instead, a representative configuration is chosen. The results of the TVAC tests are used for a heat transfer model to derive a rough value for the overall, average surface emissivity of the outer spacecraft panels. A simplified thermal balance model is used to estimate the thermal behaviour of the S-Band and X-Band antenna patches, because these patches are elevated and thermally isolated from the surface of the spacecraft outer panels. Finally, collected on-orbit telemetry is presented and compared to the TVAC test results, where possible. This on-orbit data allows a first insight into the influence of the high-powered S-Band transmitter, which are not activated during the TVAC tests, to avoid damaging the power amplifiers. The data shows that the significant transmitter power of 10 watts results in a sharp increase of the S-Band temperature itself but has little impact on the overall system temperature, due to its short time of operation.

II. OPS-SAT Thermal Design and Tests

OPS-SAT is a tightly packed spacecraft with little room for additional thermal measures. Except for some components developed by TUG, all modules are mounted inside the ClydeSpace 3U structure without any thermal modifications. The individual modules are stacked on top of each other, guided by four Titanium rods at the long edges of the structure and interconnected by PC/104 connectors. The mechanical connection between the rods and the PCBs is loose, without any additional thermal conductor such as thermal paste. Additional wire harness was used where necessary and may make minor contributions to heat distribution. The CubeSat structure is enclosed by four cover plates, on top of which the solar arrays and body panels are mounted. An outline of the panels is shown in Fig. 3. The solar arrays are held in place by screws and are in mechanical contact with the cover plates of the structure. To decrease the thermal absorptance to emissiviy ratio and to improve the heat dissipation on the surface, the umbilical/S-Band panels and the S-Band and X-Band patch antennas are covered with white AZ93 paint. The 2U umbilical panel is not a regular FR4 PCB but a special insulated metal substrate (IMS), to increase the heat transfer from the structure towards the radiating white surface of this panel.

There are some highly power-consuming payloads onboard OPS-SAT, that will be addressed in more detail below. Noteworthy are the S-Band transceiver and the X-Band transmitter, both Commercial Off The Shelf (COTS) components, manufactured by Syrlinks. Each of those peaks out at around 10 watts of power consumption, if the transmitter is turned on. The S-Band transceiver is used as default communication unit on OPS-SAT and is designed for a maximum continuous operation of 15 minutes, i.e. during a ground station pass, plus some margin. The S-Band transceiver had to be kept in idle mode during the TVAC test campaign, resulting in a minor contribution to the overall power budget of around 1.3 to 1.4 watts. As such, its full contribution to the thermal budget could not be evaluated during the tests. The X-Band transmitter was powered off during the TVAC tests.

The core of OPS-SAT experiments are the two redundant SEPPs. A SEPP is typically powered on during active experiments, or tests and maintenance by the ground control team. Each SEPP‘s power range lies in between 3.5 watts and 8 watts, that is HPS and FPGA parts combined. To avoid overheating, additional thermal measures are applied to ensure proper cooling and heat distribution. The so-called SEPP stack is a combination of the two SEPP boards, the SDR board and a magnetorquer board, encapsulated by an anodized aluminium housing. Thermal gap filler material is placed to thermally link the SEPP System On Chip SOC package with the PCB above and below, to improve heat transfer away from the SOC module. The gap filler is extended towards the edges of the PCB, all the way to contact the aluminium housing. As such, the surrounding PCBs and the housing act as a heat sink. Further heat transfer is achieved by individual frame elements, that thermally couple the outer 2 millimetres of the PCB edges of all SEPP stack PCBs to the housing. This thermal link is further improved by application of thermal paste. The design of the SEPP stack housing improves the thermal link between the enclosed PCBs and the CubeSat structure by providing a larger contact surface area towards the edge-mounted rails of the structure. This design improves heat transfer towards the rails and the outer panels.

OPS-SAT is fully covered by solar panels, except for the S-Band patches, the umbilical panel and the bottom panel. The umbilical panel is a metal PCB painted in white. Unfortunately, the thermo-optical properties of this white paint are not provided by the manufacturer. The S-Band antennas are custom-designed by TUG and are right hand circular polarized patch antennas. The copper patch is held in place by an eight millimetre strong Rohacell layer and can be considered thermally well isolated [2] from the rest of the spacecraft, except for the small antenna feed wire. So it is crucial that the thermo-optical properties of the copper patch are adjusted by thermal coating to stay within a reasonable temperature range, when in sunlight. White AZ93 paint was chosen as cover material for the patch antennas, providing an α/ε of 0.165 [3]. Without the paint, the antennas could reach destructively high temperatures, if the patch is viewed as thermally isolated from the rest of the spacecraft. An example of this extreme case is given in section IV.A.

A. Thermal Vacuum Tests

The OPS-SAT TVAC tests have been carried out at the facilities of RUAG Space in Vienna. The tests are performed with the OPS-SAT protoflight model [4] and followed an approach of demonstrating functionality under uniform environmental conditions. As such, the spacecraft is exposed to a uniform temperature on all sides during the tests. To ensure safety of the most critical units, these were equipped with additional temperature sensors during the tests. The battery of the Electric Power Supply (EPS) is the most temperature critical unit and as such, it is chosen as Temperature Reference Point (TRP). As the precise thermal behaviour of the spacecraft was unknown at the beginning of the tests, it was decided to start with a thermal balance phase. This thermal balance phase is used to identify the temperatures at all sensors with respect to the TVAC chamber temperature, and the individual subsystems of the spacecraft, once the spacecraft is thermally stabilised. As the nature of OPS-SAT is being used for various types of experiments, an example case was chosen to demonstrate a possible medium power load case. Noteworthy for this load case is the SEPP which

consumes roughly 5.5 watts as a single unit. The power output throughout the spacecraft is roughly 12.14 watts. The UHF

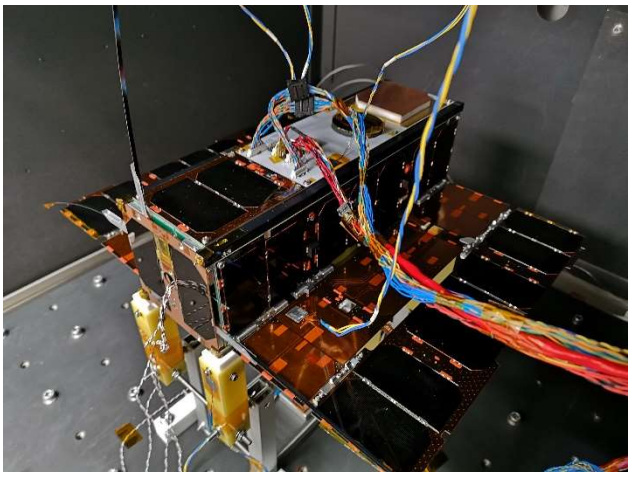


Fig. 1 OPS-SAT in TVAC chamber, resting on the Mechanical Ground Support Equipment (MGSE). Thermal sensors and umbilical harness are connected.

and S-Band radio transmitters could not be powered during the TVAC tests and as such, the effects of an additional 9 watts of power – compared to idle mode - cannot be demonstrated during the TVAC tests. However, the heating of the S-Band transceiver is shown in the on-orbit results, in Section V.

The total power of 12.14 watts is mostly dissipated through the eight body-mounted panels and the 3U structure. The deployable solar wings only provide a minor contribution to heat dissipation since the hinges on which they are mounted provide a very weak thermal link towards the spacecraft structure. The eight body-mounted panels are shown in Fig. 3 and comprise of the solar arrays, the S-Band antenna panels and the auxiliary bottom panels. Each solar array is equipped with a temperature sensor, mounted on the space facing side of the PCB, with the exception of the 1U panel on top of the spacecraft. An additional temperature sensor is mounted on the outside of the bottom auxiliary panel. During the TVAC tests, additional external temperature sensors have been placed on the inside and the outside of the spacecraft. Two sensors are placed on each of the diagonally opposite structure rails, at roughly a distance of 1U from the top and bottom respectively. One sensor is placed on the 2U umbilical / S-Band panel, located on the +X side of the spacecraft, to have a comparison to the sensor on the 1U solar array, also located on the +X side.

The sensors and surfaces used for the heat transfer model are shown in Fig. 3. The x-sides are highlighted in red, the y-sides in green and the z-sides in blue. The labelled arrows indicate the location of the panel-mounted temperature sensors. The sensors, TP4 to TP8, as well as the +Z sensor, are indicated as a red encircled cross. TP4 to TP8 are externally mounted sensors and are not part of the spacecraft. The surfaces are labelled in the bottom left corner of each highlighted area and the remaining grey surface indicates the visible part of the structure.

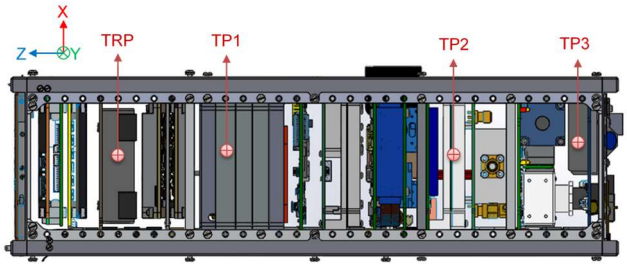


Fig. 2 Temperature sensors on the battery (TRP), SEPP (TP1), S-Band transceiver (TP2) and Optical receiver housing (TP3).

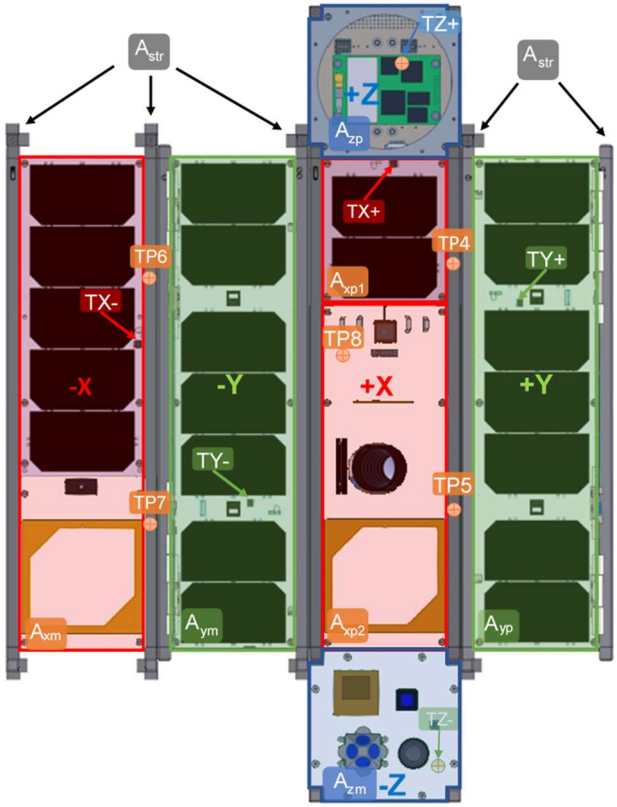


Fig. 3 Temperature sensors used and surfaces regarded for the heat transfer model. The +X surface is split into two parts, A_{xp1} and A_{xp2}. Sensors TP4 to TP8 are externally mounted during the tests.

III. Thermal Balance Phase

The thermal balance phase is crucial to understand the temperature distribution throughout the OPS-SAT protoflight model so that no unit is operated above its allowed operational temperature limit. In order to ensure that no limits are reached already during this phase, a low TVAC chamber temperature of 5 degree Celsius was chosen as a starting point. Fig. 4 shows the temperatures at the individual sensors during the heat-up and steady-state periods of the thermal balance phase. It shows that the 5 degree Celsius ambient temperature was already close to the maximum allowed ambient temperatures for the test, bringing the battery (TRP) close to its operational maximum of 45 degree Celsius. The dual sensors on the

structure show a temperature gradient, with the higher temperatures close to the SEPP.

The top panel of Fig. 4 shows nicely that the temperatures of the SEPP and the battery lie close to each other, with the battery slightly following behind the increases of the SEPP. The jumps in SEPP temperature are tied to different power levels used during the tests. The position of sensors TRP, TP1, TP2 and TP3 was chosen to get more or less equidistant temperature measurements throughout the height of the satellite. The temperature drop towards TP2 and TP3 nicely shows the distance to the more power consuming SEPP.

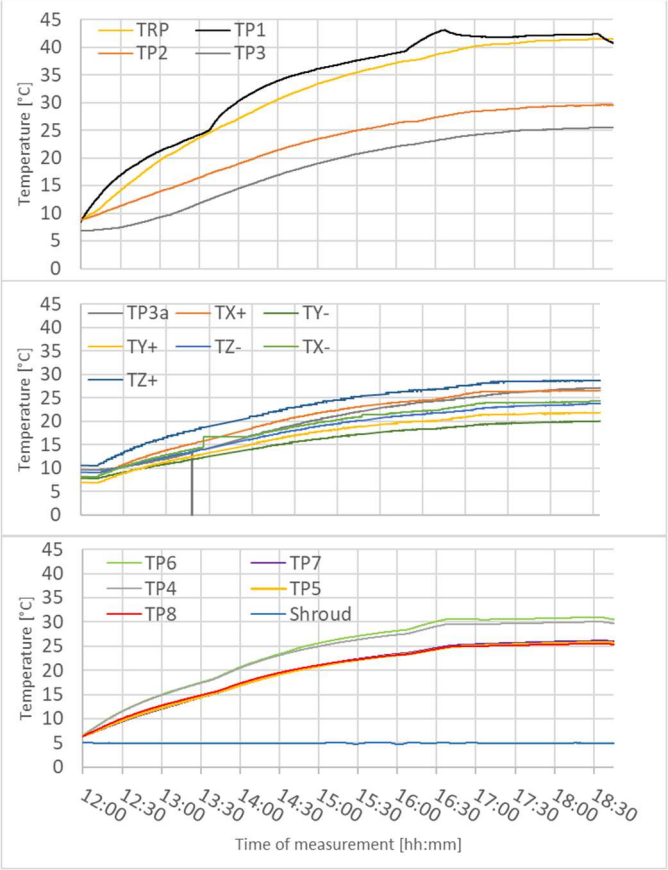


Fig. 4 TVAC test thermal sensor measurements. The additionally added temperature sensors are shown in the top graph: Battery (TRP), SEPP (TP1), S-Band transceiver (TP2) and optical receiver housing (TP3). The middle graph shows the temperatures of the outer surfaces of the spacecraft and the bottom graph shows the sensors placed on the structure (TP4 to TP7), the sensor on the X- panel (TP8) and the TVAC chamber shroud.

The middle panel shows the temperatures of the spacecraft surfaces and an additional sensor inside the optical receiver housing. This sensor is labelled as TP3a as it sits close to the sensor TP3.

Variations in temperature on the individual surfaces is a result of the varying thermal links between the heat-producing units to the respective surfaces. Another contribution to variations in readings is the positioning of the temperature sensors on the panels, some of them closer to the heat-producing units than others.

The additional external sensors are shown in the bottom panel of Fig. 4. TP4 and TP6 are the two sensors closest to the SEPP and show higher temperatures. TP5 and TP7 are located further down the structure and show lower temperatures. TP8 is located on the X+ panel and also show low temperature, slightly less than the 1U solar array on the X+ side. The temperature of the TVAC chamber shroud is shown in blue, and was kept constant at 5 degree Celsius during this period.

IV. Heat Transfer Model

By making use of the thermal measurements during the TVAC tests, an attempt is made to obtain the mostly unknown thermal emissivity of the OPS-SAT surfaces. Once the thermal emissivity is known, the distribution of thermal waste heat inside the spacecraft is derived. The advantage of this approach is the complete elimination of an underlying thermal model. As such, no assumptions have to be made about the thermal links and the thermal properties of the materials that link the heat-producing units to the surfaces of the spacecraft. Since the TVAC measurements yield the results of the overall effects created by the internal behaviour of the spacecraft, the power fraction of the total electrical power that arrives at each surface can be calculated directly.

Table 1 Surfaces equipped with temperature sensors during TVAC tests

Name	Area [$*10^{-3}m^2$]	Temperature [°C]	Description
A_{str}	24.1	26.72	Visible part of CubeSat structure
A_{xp1}	8.1	26.80	1U solar panel on X+ side
A_{xp2}	14.9 / 18.8	24.42	2U panel on X+ side without/with S-Band antenna
A_{xm}	23 / 26.9	24.30	2U solar panel on X- side without/with S-Band antenna
A_{yp}	26.9	22.00	3U solar panel on Y+ side
A_{ym}	26.9	20.10	3U solar panel on Y- side
A_{zp}	9.3	28.73	1U solar panel on Z+ side
A_{zm}	9.6	23.80	1U auxiliary panel on Z- side

Following the principle of conservation of energy, the total electrical power consumed by the spacecraft will dissipate as heat throughout the total surface area of the spacecraft. The total surface area is split into eight parts, representing the eight distinguishable temperatures derived from the surface of the spacecraft during the TVAC tests. Table 1 shows the model input parameters, the surface area (see Fig. 3) and the corresponding surface temperature. The temperature of the structure set to the average of the four sensors TP4, TP5, TP6 and TP7.

The structure is constructed of aluminum, with its surface anodized in black. The infrared emissivity, ε, of the structure surface is chosen as 0.834-0.856 [5]. The remaining surfaces mainly comprise of solar cells, PCB material and white paint, all with an infrared emissivity in the same order of magnitude [6], [7], [8], [14]. Therefore, the emissivity of the other surfaces is assumed to be equal and can be calculated with the help of the heat flux equation [9]:

$$Q = \varepsilon\sigma(T^4 - T_{amb}^4) \quad (1)$$

With Q = P / A, where Q is the electrical power, P, that is dissipated at the satellite surface, and A is the satellite surface. The Boltzmann constant is notated as σ. T_s is the surface temperature of the heat-emitting surface and T_{amb} is

the ambient temperature; in this case, the temperature of the TVAC chamber shroud. Considering all individual surfaces, the total heat flux is the sum of the heat flux on all individual surfaces [10]. The heat flux at the structure is known by choosing an average thermal emissivity from [5], ε = 0.845. With this, it is possible to separate the right-hand part of (1) into two parts, one with known emissivity and one with unknown emissivity:

$$P = \varepsilon_{str}\sigma A_{str}(T_{str}^4 - T_{amb}^4) + \varepsilon_{av}\sigma \sum_{i=2}^{N} A_i\left(T_i^4 - T_{amb}^4\right) \qquad (2)$$

ε_{str} is the known emissivity of the structure surface area and ε_{av} is the unknown average emissivity of the remaining surfaces. Equation (2) is used to derive the average emissivity of the satellite panels:

$$\varepsilon_{av} = \frac{P - \varepsilon_{str}\sigma A_{str}(T_{str}^4 - T_{amb}^4)}{\sigma \sum_{i=2}^{N} A_i(T_i^4 - T_{amb}^4)} \qquad (3)$$

Now that the unknown emissivity parameter is set, (1) is used to obtain the power fraction that is dissipated at each panel:

$$P_i = \varepsilon_i \sigma A_i (T_i^4 - T_{amb}^4) \qquad (4)$$

P_i is the fraction of the total consumed power that arrives at the individual panel surfaces A_i. The thermal emssivity, ε_i, is set to ε_{str} for the structure and to ε_{av} for all remaining surfaces. The power fraction is particularly useful for temperature estimations under various orbital scenarios and attitude conditions.

Since the exact thermal behaviour of the S-Band patches is unknown, the calculations are performed for two cases. Case (a) considers the S-Band patch surface area as thermally completely isolated and as such, this surface does not contribute to thermal radiation. Case (b) treats the S-Band patches as a regular surface, fully contributing to thermal radiation. The average emissivity, calculated with the help of (3) is 0.84 for case (a) and 0.79 for case (b). The results for both cases are shown in TABLE 2. The surface of the patch antennas is part of the surface area of A_{xp2} and A_{xm} respectively. This is directly reflected in the power dissipated at these panels. Case (b) shows the increased power fraction at P_{xp2} and P_{xm} due to an increased radiation surface area, compared to case (a). As such, all other panels must show a decreased power fraction for case (b).

TABLE 2 DISSIPATED POWER AT EACH PANEL, WITHOUT AND WITH PATCH ANTENNAS

Name	Power (a) [W]	%	Power (b) [W]	%
P_{str}	2.4319	19.74	2.4319	19.74
P_{xp1}	0.8110	6.15	0.7581	6.58
P_{xp2}	1.3237	12.67	1.5609	10.75
P_{xm}	2.0233	17..96	2.2127	16.43
P_{yp}	2.0600	15.63	1.9257	16.72
P_{ym}	1.8117	13.75	1.6936	14.71
P_{zp}	1.0319	7.83	0.9646	8.38
P_{zm}	0.8243	6.26	0.7705	6.69
Sum	**12.1378**	**100**	**12.1378**	**100**

A. Patch Antennas

The temperature of an isolated surface illuminated by the Sun can be estimated based only on the incoming solar flux and the thermo-optical properties of the surface [11].

$$T = \sqrt[4]{\left(\frac{S}{\sigma}\frac{\alpha}{\varepsilon}\right)} \qquad (5)$$

Where T is the temperature of the surface in Kelvin, S is the solar flux in watts per square metre, α is the solar absorptance coefficient, σ is the Stefan-Boltzmann constant and ε is the infrared emissivity coefficient. The OPS-SAT antenna patches are made out of copper and we use the thermo-optical properties range for copper with α = 0.32 to 0.55 and ε = 0.02 to 0.04 [7]. The thermal paint used on the antennas is AZ technologies white paint AZ-93. For calculating the temperature, a solar flux of 1360.8 watts is assumed [12]. This yields a temperature range of 354.4 to 512.9 degree Celsius, which would be too hot for the underlying Rohacell layer and induce significant thermal noise on the antenna. Treating the antenna patch with AZ-93 white paint leads to a significantly lower temperature of -24 degree Celsius under the same conditions. Since there is no perfect isolation, the will be weakly thermally linked to the spacecraft and the patch temperature can be expected to be above -24 degrees Celsius.

V. ON-ORBIT TELEMETRY

Fig. 5 shows telemetry collected at a pass during the OPS-SAT commissioning phase. The top graph shows the temperatures of the EPS battery as blue line and the temperatures at the EPS mainboard, Array Conditioning Units (ACUs) and Power Distribution Units (PDUs). For this particular pass, the battery is close to temperature observed during the thermal balance phase at the TVAC tests, shown in the top graph of Fig. 4. The EPS mainboard, ACUs and PDUs are several degrees higher as these are active components.

The middle graph shows the temperatures of the OPS-SAT OBC and the UHF transceiver, all in the same range as the EPS components. It has to be noted that the OBC and UHF transceiver are also located right next to the EPS, at the very top of the spacecraft, next to the Z+ side.

The bottom graph shows the temperature of the S-Band transceiver (compare to TP2 in Fig. 2) in idle mode at the beginning of the pass and with active transmitter towards the last third of the pass. The rise in temperature with active transmitter is clearly visible in the "S-Band PA" curve (green). In addition, the bottom graph shows the temperatures recorded at the body panels of the spacecraft (see middle graph of Fig. 3 for comparison). A wider span of temperatures on the body panels is apparent, since the environmental conditions in orbit are not uniform as during TVAC testing. In particular, the temperature of the Sun facing X- panel is at a round 65 degree Celsius. While the other panels are in a temperature range that is comparable to the observations during TVAC testing.

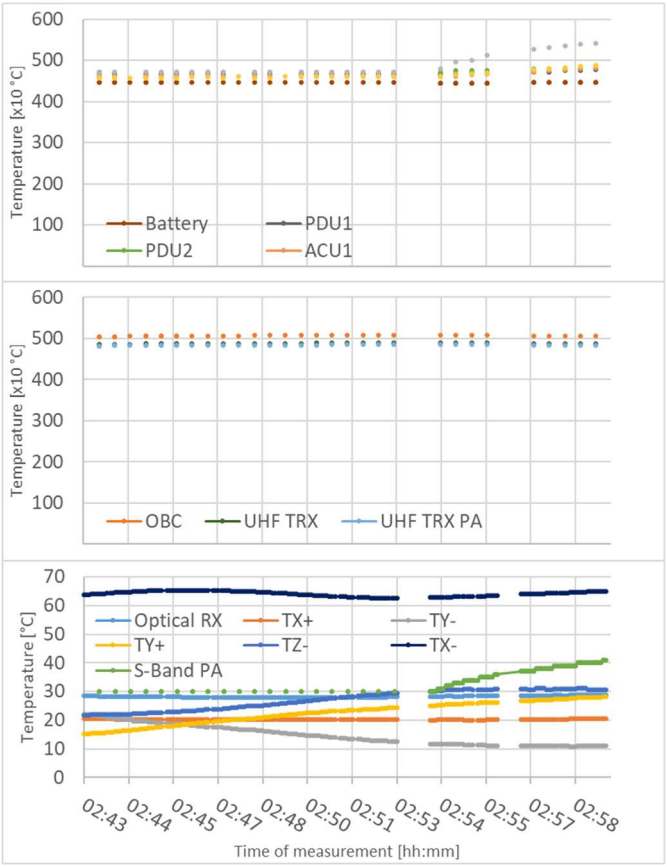


Fig. 5 On-orbit temperature telemetry of OPS-SAT EPS (top), OPS-SAT bus (middle) and body panels plus S-Band transceiver (bottom). Temperatures for the top two graphs show at a 0.1°C scale, and the bottom graph shows 1°C scale. Note the different range and scaling of the bottom graph.

The influence on the S-Band transmitter on the overall system and battery temperature is observed to be marginal, despite the high power output of up to 10 watts. It is designed for a maximum continuous operation of 15 minutes and under nominal operations, the S-Band transmitter is only turned on for several minutes, depending on pass duration and conditions. As such, the total amount of thermal energy brought into the spacecraft is rather low.

It has to be noted that the data in Fig. 5 is recorded at a different payload configuration as during the TVAC test. In particular, the SEPP was powered at this point. At this period, the spacecraft was facing the Sun with its X- side, which means that the on-orbit conditions under full sunlight can be considered hotter than during the TVAC thermal balance phase.

Fig. 6 shows the telemetry recorded at two passes during OPS-SAT commissioning. In the first pass on the left, the S-Band transmitter is powered on for 5 minutes and 35 seconds, during which the rise in temperature is 11 degrees Celsius. During the second pass, the temperature rise is observed at 10 degrees Celsius during a 6 minute and 20 second period, which yields a rise of 1.7 to 2.0 degree per minute of transmitter operation. The rise in battery temperature is comparably low at 0.2 to 0.3 degree Celsius during both passes. A similar behaviour is observed at the panel temperatures, shown in the bottom graph of Fig. 5. There are two main reasons for this behaviour. First, the battery and the panels are at a distance to the S-Band transceiver and second the overall heat-capacity of the spacecraft. This reflects also in the delayed heating of the battery and panels.

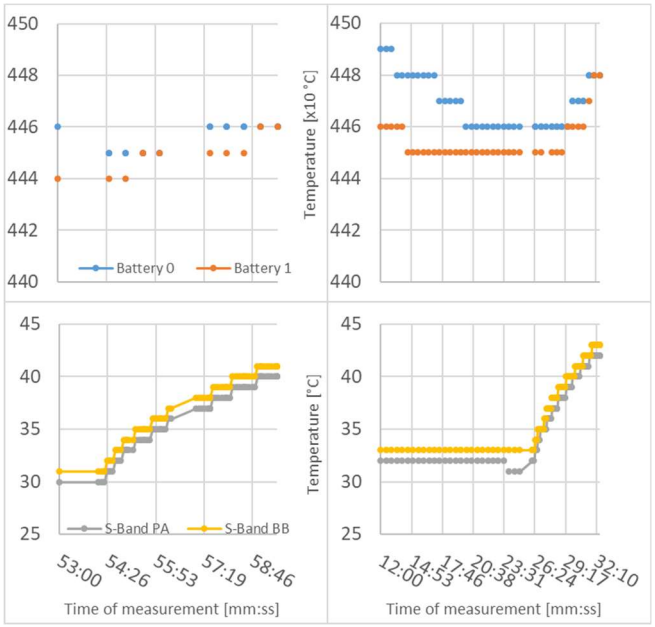


Fig. 6 Battery temperature (top) and S-Band transmitter temperature (bottom) at two passes during OPS-SAT LEOP. Note that the battery temperatures are shown at a different temperature range of degree, with a scaling factor of 10.

VI. Discussion

The results of the OPS-SAT TVAC tests shown in this work, represent only a fraction of the actually conducted tests. The crucial thermal balance phase is chosen as an example for the heat-up and steady state phase, as it reveals the temperature relations in between the powered components. The results of the heat transfer model derived from the TVAC measurements are to be viewed with a good portion of caution, as several assumptions are made.

First, the temperature is assumed to be uniform on each surface, whereas in reality, there will be a gradient due to non-uniform heat sources. This behaviour is confirmed by the temperature gradient observed at the sensors TP4 to TP7, placed on the structure rails.

Second, the temperature sensors are located at varying distances from the heat sources, which cannot be taken into account by the model in its current form. Third, the contribution of the S-Band patch antennas to heat dissipation is unknown, therefore the two presented cases (a) and (b) (see Section IV) indicated the range of uncertainty. As such, the heat transfer model is intended to provide an estimate of the order of magnitude of the heat dissipated by each body panel and the CubeSat structure and not as a replacement for a comprehensive thermal model. The advantage of the approach presented here, is the simply analytical solution, obtained by using measured data as an input. No detailed knowledge or assumptions about the thermal links and junctions inside the spacecraft are required. The overall thermal power fraction that arrives at the individual

spacecraft surfaces can directly be calculated. Future work will include the correlation of the heat transfer model results to on-orbit data, to verify the validity of the model. Parts of the already available on-orbit data is presented to add an insight on the thermal contribution of the power consuming S-Band transmitter, which could not be powered on during the TVAC test campaign. Additionally, the insights from the OPS-SAT mission will provide valuable information for the upcoming PRETTY mission (unpublished [13]), also a 3U CubeSat.

ACKNOWLEDGMENT

The work described in this paper is being carried out under ESA Contract 4000117431/16/D/SR. The authors are grateful to Mr. David Evans, the Technical Officer at ESOC and all his colleagues at ESOC for the continuing support. The authors also wish to acknowledge the support by the Austrian aeronautics and Space Agency, in particular Dr. Stephan Mayer.

REFERENCES

[1] D. Evans, O. Koudelka, L. Alminde and K. Schilling, “The ESA OPS-SAT CubeSat Mission,” Majorca Island, Spain, 2014.

[2] M. Flori, V. Puţan and L. Vîlceanu, “Using the heat flow plate method for determining thermal conductivity of building materials,” in *IOP Conf. Ser.: Mater. Sci. Eng. 163 012018*, 2016.

[3] “AZ-93 White Thermal Control, Electrically Conductive Paint / Coating,” AZ Technologies, [Online]. Available: http://www.aztechnology.com/materials-coatings-az-93.html.

[4] “CubeSat.org,” 2020. [Online]. Available: https://static1.squarespace.com/static/5418c831e4b0fa4ecac1bacd/t/56e9b62337013b6c063a655a/1458157095454/cds_rev13_final2.pdf. [Accessed 27 04 2020].

[5] A. Gustavesen and P. Berdahl, “Spectral Emissivity of Anodized Aluminum and the Thermal,” *Nordic Journal of Building Physics,* 2003.

[6] M. M. Finckenor and R. F. Coker, *Optical Properties of Nanosatellite Hardware,* NASA/TM-2014-218195, 2014.

[7] L. Kauder, *Spacecraft Thermal Control Coatings References,* NASA/TP-2005-212792, 2005.

[8] A. Riverola, A. Mellor, D. Alonso Alvarez, L. Ferre Llin, I. Guarracino, C. Markides, D. Paul, D. Chemisana and N. Ekins-Daukes, “Mid-infrared emissivity of crystalline silicon solar cells,” *Solar Energy Materials and Solar Cells, 174, pp. 607-615,* 2018.

[9] A. Raslan, G. Michna and M. Ciarcià, “Thermal Simulation of a CubeSat,” in *IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 2019, pp 453-459.*, 2019.

[10] L. Klobučar, I. Tiselj and B. Končar, *Thermal radiation heat transfer between surfaces,* Ljubljana, 2016.

[11] J. Doenecke, Thermalkontrolle von Raumfahrzeugen, 1988/2014.

[12] K. G. and J. L. Lean, “A new, lower value of total solar irradiance: Evidence and climate significance,” *Geophysical Research Letters,* 2011.

[13] R. Zeif, A. Hörmer, M. Kubicka, M. Henkel, O. Koudelka, “ A GPS Patch Antenna Array for the ESA PRETTY Nanosatellite Mission”, 2020

[14] A. Riverola, A. Mellor, D. A. Alvarez, L. F. Llin, I. Guarracino, C. N. Markides, D. Paul, D. Chemisana and N. Ekins-Daukes, “Experimental and theoretical study of the infrared emissivity of crystalline solicon solar cells,” in *IEEE 44th Photovoltaic Specialist Conference (PVSC), Washington, DC, 2017, pp. 1339-1341.*, 2017.

From OPS-SAT to PRETTY Mission: A Second Generation Software Defined Radio Transceiver for Passive Reflectometry

Reinhard Zeif
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
reinhard.zeif@tugraz.at

Andreas Hörmer
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
hoermer@tugraz.at

Manuel Kubicka
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
manuel.kubicka@tugraz.at

Maximilian Henkel
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
henkel@tugraz.at

Otto Koudelka
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
koudelka@tugraz.at

***Abstract*—After the successful launch of the ESA OPS-SAT Nanosatellite in December 2019, the Institute of Communication Networks at Graz University of Technology (TUG) has started its work on a new second-generation Software Defined Radio (SDR) transceiver platform for the ESA PRETTY mission. The mission goal of PRETTY is the demonstration of the passive reflectometry concept with an SDR on a 3U Nanosatellite. The PRETTY satellite requires a powerful second-generation SDR receiver that extends the functionality and performance of the first-generation SDR used for OPS-SAT. There are many lessons learned about the first-generation SDR characteristics, the performance, ease of use and the strengths but also the weaknesses of the design during the OPS-SAT environmental and functional testing campaign. The second-generation SDR design considers the experiences from the first-generation SDR and implements several improvements for the thermal behavior, mechanical sustainability, device control and status monitoring in order to achieve higher overall performance and reliability. The second-generation SDR uses an AD9361 radio frequency (RF) frontend chip, that allows the signal reception with two independent receive channels and signal transmission with two independent transmit channels. In particular, the new transmit functionality of the second-generation SDR is a remarkable improvement compared to the first-generation SDR for OPS-SAT, due to its full-duplex, bidirectional communication capabilities. Further improvements provide the possibility, to extend the design with RF mixer boards, to achieve the flexibility required for future applications on higher RF bands.**

Keywords—PRETTY, Nanosatellite, Software Defined Radio, RF front end, Passive Reflectometry

I. Introduction

The first generation of the SDR system for Nanosatellites was developed by Graz University of Technology (TUG) in cooperation with MEW Aerospace for the ESA OPS-SAT mission in the years 2014 to 2019. The OPS-SAT first-generation SDR frontend is a simple receiver design, based on a LMS6002D RF frontend chip, with the purpose to extend the so-called Satellite Experimenters Processing Platform (SEPP) with an RF receive functionality. The SEPP is the heart of the OPS-SAT payloads. The SEPP executes all experiments and performs the required baseband data processing for the first-generation SDR receiver. The SEPP additionally provides the interfaces to the satellite ground station allow the uplink of Telecommands (TCs) to the SDR and SEPP. After successful launch of the ESA OPS-SAT Nanosatellite in December 2019, the Institute of Communication Networks and Satellite Communications (IKS) at TUG has started its work on a new second-generation SDR transceiver platform for the ESA PRETTY mission.

The goal of the ESA PRETTY mission is to demonstrate the concept of passive reflectometry with a cost-efficient 3U Nanosatellite. Martín-Neira first proposed this method in 1993 [1] and Lowe et al. tested the method successfully in 2002 [2]. Passive reflectometry describes a process that uses reflected Radio Frequency (RF) signals in combination with signal correlation and processing for the characterisation of planetary surface properties and height [3]. In contrast to active reflectometry, passive reflectometry does not actively generate and transmit RF signals, but rather uses existing RF signals from other external sources such as Global Navigation System Satellites (GNSS). The general concept of passive reflectometry is shown in Figure 1.

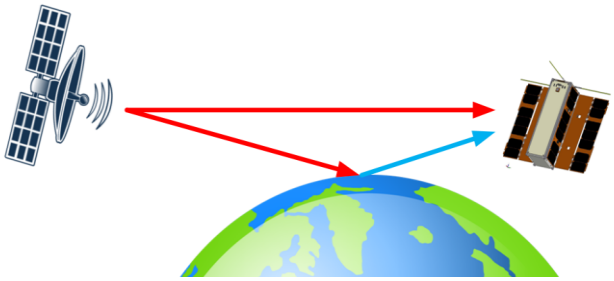


Fig. 1. The concept of passive reflectometry using a direct and reflected signal from an external signal source.

The passive reflectometer is implemented as Software Defined Radio (SDR) platform, consisting of an RF frontend with an external antenna and a baseband data processing unit. The IKS at TUG is responsible for the spacecraft system design and the development of the second-generation SDR that is used for the PRETTY passive reflectometry system.

The PRETTY mission imposes strict requirements for the passive reflectometer regarding the RF frontend performance, sampling rate and total signal gain. These requirements are beyond the capabilities of the first-generation SDR, so that it is necessary to develop a new, second-generation SDR system. Beyond that, the second-generation SDR allows the transmission of RF signals so that it can be used as general-purpose transceiver in future missions.

Chapter II describes the general design of the first-generation SDR with its strengths and weaknesses regarding architectural, software, thermal and mechanical design. Chapter III discusses the improvements derived from the first-generation SDR and the final design of the second-generation SDR.

II. The First-Generation SDR for OPS-SAT

A. Overview

The first-generation SDR receiver was developed for the OPS-SAT Nanosatellite mission as hardware extension of the Satellite Experimenters Processing Platform (SEPP).

The OPS-SAT Nanosatellite is a test laboratory in space that demonstrates new operations concepts and state-of-the-art hardware and software technologies [4]. Based on the mission requirements, the first-generation SDR is designed as an experimental payload of opportunity that allows the reception of RF signals between 300MHz and 1.5GHz.

It was decided by the project to use the Lime Microsystems MyriadRF radio frontend Printed Circuit Board (PCB) in combination with a custom SDR motherboard to reduce the costs and development time. The MyriadRF frontend is a Component off-the-Shelf (COTS) development kit board for the LMS6002D RF frontend (RFFE) chip, with a transmit (TX) power amplifier, a receive (RX) low noise amplifier (LNA) and matching components. The motherboard extends the functionality of the MyriadRF, to make it compatible with the OPS-SAT system design and the SEPP baseband processing hardware.

Figure 2 shows the MyriadRF PCB with the LMS6002D chip on its top side and a mezzanine connector for the connection to a motherboard on its bottom side. Two Sub-Miniature-A (SMA) connectors act as the RX signal input and TX signal output.

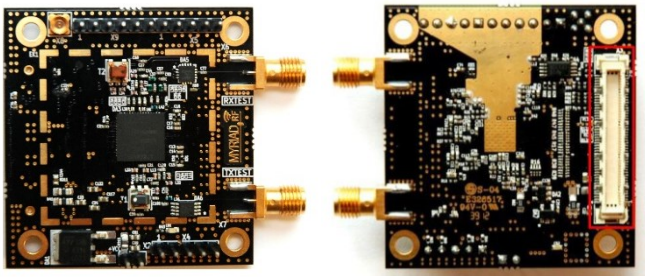


Fig. 2. The MyriadRF frontend module top (left) and bottom side (right). A mezzanine connector on the bottom side connects the MyriadRf to a motherboard that includes peripheral components and a baseband signal processing unit [5].

B. The LMS6002D RF Frontend

The LMS6002D RF front-end chip covers a frequency range from 0.3 to 3.8 GHz and consists of an RX and a TX channel. The utilization of the RX and TX channels allows bidirectional full-duplex communication with a maximum channel bandwidth of 28 MHz.

Figure 3 shows the internal architecture of the LMS6002D chip.

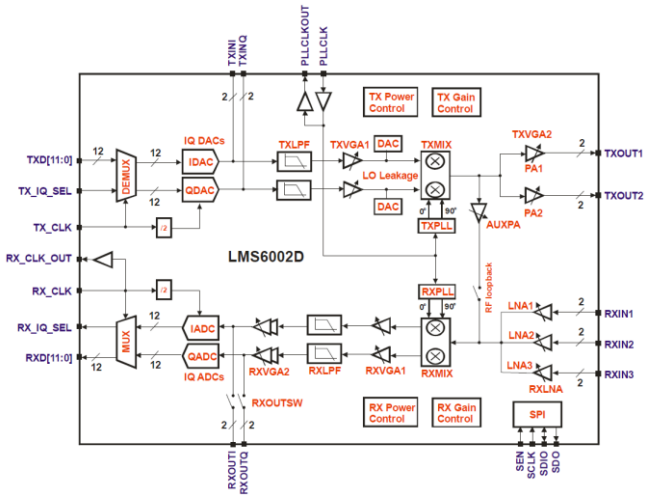


Fig. 3. The LMS6002D architectural block diagram. The chip consists of two quadrature modulators for signal transmission and reception and uses a SPI interface for device configuration and monitoring [6].

The LMS6002D RX channel consists of a quadrature demodulator to downconvert the RX input signal to the baseband. . Three RX inputs allow the switchover between different external RF circuits, if an optimized circuitry for different frequency bands is required. An LNA together with two Variable Gain Amplifiers (VGAs) allow the amplification of the internal differential baseband signals before and after demodulation. A baseband Low-Pass Filter (LPF) restricts the baseband signal spectrum before sampling with the Analog-to-Digital Converter (ADC), to avoid aliasing. A Parallel I/O bus transmits the resulting IQ samples to a baseband processing unit

The LMS6002D TX channel contains equivalent components but uses a modulator to upconvert the baseband IQ samples to the transmitted RF signal.

The LMS6002D chip architecture does not include functionalities for hardware-level Automatic Gain Control (AGC) and Automatic Frequency Control (AFC). Functionalities of that kind have to be implemented inside the baseband processor with the consequence of an increased development time and verification effort.

The LMS6002D requires an external controller that implements particular setup and calibration algorithms for the LNA, VGA and Phase-Locked Loop (PLL) blocks, whenever the chip is powered on or some gain setting is changed. This external controller is implemented in the SEPP and the SDR firmware. The firmware executes the SPI read and write transaction to the LMS6002D registers. The calibration and configuration algorithms require a “read-modify-write-verify” approach and use a loop-based design to guarantee the validity and consistency of the applied low-level register settings. The consequence of the algorithms is a large number of SPI read and write transactions and results in a very high temporary SPI bus workload.

It has to be emphasized that it was challenging to define an appropriate set of configurations for the LMS6002D registers during the OPS-SAT project because of the limited

documentation and the expenditure of time to verify the functionality of the LMS6002D signal chain. It turned out that simple tasks like the configuration of the overall receiver gain are very challenging because. For example, it can happen that an arbitrary block in the signal chain does not behave as expected, but it is not detectable which one.

It is only possible to solve such problems, by checking all signal path blocks and analyze their configuration values to identify the cause of the problem.

From the experiences made with the first-generation SDR, we conclude that it is better to use a more advanced RF frontend design for the PRETTY second-generation SDR with an RF frontend chip that provides more information on the internal chip status. The second-generation SDR frontend chip shall not require all necessary algorithms to be implemented by hand. Instead, critical algorithms shall be part of the chip so that the configuration is more user friendly. This reduces the development time, decreases the SPI bus load and allows to use third-party software for signal processing and packet-based communication.

C. *System Architecture*

As shown in Figure 2 and 4, the MyriadRF is a daughter board and not a fully functional stand-alone device. The MyriadRF daughter board has to be mounted onto a motherboard PCB that contains a clock generator for the LMS6002D, a temperature sensor and some protection hardware for the interfaces to the SEPP baseband data processing unit. When mentioning the SDR, this always refers to the motherboard with the MyriadRF daughter board and the SEPP baseband processing unit as a whole.

Figure 4 shows the final hardware architecture of the SDR RF frontend with a MyriadRF daughter board and some peripheral components.

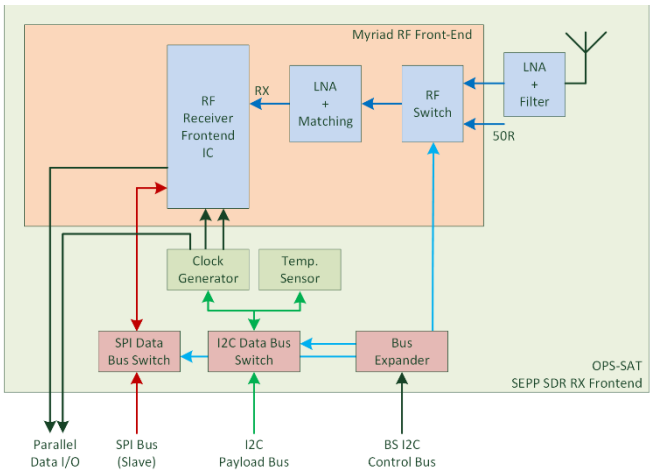


Fig. 4. The architecture of the SDR receiver RF frontend with a MyriadRF daughter board. The design allows the reception of RF signals with a single antenna only and provides no TX functionality.

The RX channel is equipped with an extra 15dB LNA and a 1.5GHz Low-Pass Filter (LPF). The filter was inserted to block any incoming high-power signals generated by the OPS-SAT on-board S-Band transmitter. A high-precision clock generator with a temperature compensated crystal oscillator (TCXO) generates a reference clock for the LMS6002D RX Phased-Locked Loop (PLL) and a sampling clock for the RX Analog-Digital Converter (ADC).

It has to be emphasized that the MyriadRF TX channel is deactivated and not available during the various OPS-SAT experiments to prevent the RF signal transmissions on protected frequency bands due to a human-fault.

The SDR motherboard contains the following components to improve the performance and fail-safety of the SDR:

- A temperature sensor to be able to switch the system off in case of thermal overheating.
- A second LNA to increase the signal power of weak signals.
- Data bus switches that allow the electrical separation of the I2C and SPI data interfaces from the satellite bus and SEPP, so that the interfaces are not fully blocked by a single faulty I2C unit.

The SDR motherboard is directly connected with the SEPP hardware that uses a SPI master and an I2C master controller to configure the MyriadRF and the motherboard before and during RF signal reception. The MyriadRF forwards the received IQ data samples on a Parallel I/O bus connected to the FPGA portion.

Figure 5 shows the interfaces between the SDR motherboard and the SEPP FPGA and HPS portion.

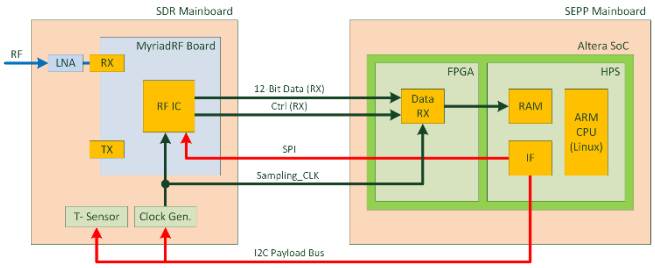


Fig. 5. The complete SDR receiver architecture consisting of the SDR motherboard with the RF frontend and the SEPP. The SEPP controls the frontend over an I2C and SPI bus. Receiver data is transferred to SEPP over a 14-bit parallel bus.

The SEPP uses an Altera Cyclone V System-on-Chip (SoC) as core component that consists of a Field Programmable Gate Array (FPGA) and a Hard-Processing System (HPS) portion. The FPGA allows high-speed signal processing in hardware whereas the HPS portion allows low-speed signal processing in software by using a Linux operating system.

D. *Software Design*

The SEPP software is a very important component of a SDR system, since most of the SDR functionalities are realized in software instead of hardware.

The SEPP software controls and monitors the SDR motherboard and MyriadRF radio frontend during experiment execution and defines the functionality of the baseband processing inside the SEPP FPGA and HPS Linux operating system (OS).

Figure 6 shows the software architecture and modules of the SEPP SDR C/C++ library API. All functionalities are organized into a package that is available to all experimenters as a C/C++ software library Application Programming Interface (API) for SEPP Linux OS. The API uses object-oriented code so that it can be reused for other devices.

The use of an SDR has many advantages compared to a classical hardware radio system because it allows the extension and update of the implemented functionality. On the other hand, even the best software is sometimes not powerful enough to overcome the disadvantages coming from an under-designed hardware.

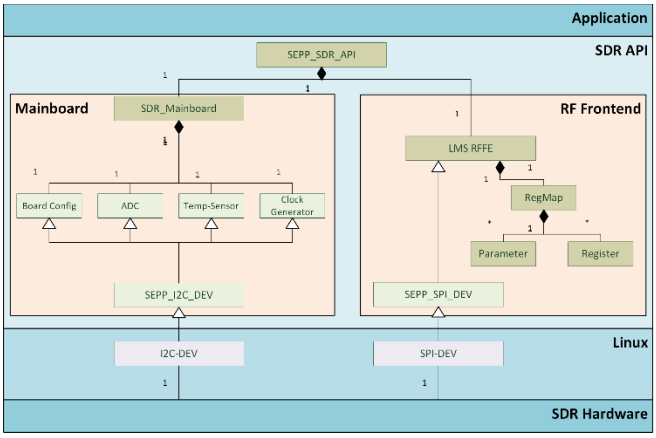


Fig. 6. The software architecture of the SDR C/C++ library API. The software consists of Linux device drivers for SPI and I2C and device-specific libraries for the motherboard components and the MyriadRF RF frontend chip.

Many required hardware features for reliable RF signal reception are not present inside the LMS6002D and have to be implemented in software. This includes the calibration but also the device control and monitoring functionalities. Especially the required low-level Phase-Locked Loop (PLL) tuning and offset calibration procedures were poorly documented so that it was necessary to do a lot of reverse engineering before we were able to implement a working software. Overall, some hundreds of methods had to be written and tested for the LMS6002D device before it was ready to be used on a Nanosatellite space mission. The unnecessary complexity of the LMS6002D could potentially be reduced significantly, by shifting the low-level hardware functions to the RF frontend chip. This would result in a reduction of the control software complexity, less design effort and lower development costs.

Because of these difficulties, we decided to use a new second-generation SDR with a more sophisticated RF frontend chip that provides better software design support.

E. Thermal Design

The SDR motherboard has an overall power consumption of about 2.5W during RF signal reception. The electrical power consumption results in a thermal dissipation inside the SDR components that increases the component temperature. Most critical are the LMS6002D with a power consumption of about 1.8 W and the external LNA on the motherboard with about 0.5 W.

The first-generation SDR is qualified for an operational temperature range of -20 to 70°C, in accordance to the nominal operational range of the satellite. The non-operational temperature range is -40 to 80°C since all LMS6002D and motherboard components are specified for this range.

For the safety of the satellite, it is important that all component temperatures lie within the operational limits, to prevent damage due to overheating. Consequently, the main goal of the SDR thermal design is to emit the required amount of thermal energy towards cold space, so that the component temperatures of the SDR and the satellite are within their operational limits. On OPS-SAT, a so-called Failure Detection, Identification and Recovery (FDIR) strategies are used to guarantee the thermal safety of the satellite components.

In vacuum, there is no gas or liquid that acts as media for the thermal convection. Hence, only the thermal conduction and the thermal radiation contribute to the thermal heat emission [7].

The rate of heat flow $\dot{Q}$ through a media is given by the following formula:

$$\dot{Q} = \lambda \cdot A \cdot \frac{\Delta T}{d} \qquad (1)$$

From (1) we can see that the rate of heat flow in materials depends on the cross-sectional surface A, the heat transfer coefficient λ, the distance d and the temperature difference ΔT between the source and sink.

Thermal energy flow and heat emission requires the presence of a thermal sink and a low thermal resistance between the source and sink so that the thermal energy of the hot source can flow to the cold sink. A good thermal sink on a Nanosatellite is typically a shadowed body-panel with a large surface that is capable to dissipate the thermal energy into space. Thermally conductive materials and an appropriate mechanical design that establishes large surface contact areas support the reduction of the thermal resistance. On PCB level, a low thermal resistance between the hot components and the PCB reduces the overall thermal resistance so that the thermal energy is transferred from the components through the PCB to the satellite thermal sink. Consequently, an appropriate thermal design uses a short distance and low resistive connection with a large contact surface between the SDR RF frontend chip, the satellite structure and a radiating body-mounted panel.

The use of a MyriadRF daughter board stands in contradiction to a good thermal design with a low resistive path between the MyriadRF and the motherboard. The connection between the MyriadRF and the motherboard consists only of four 6mm aluminium spacers, which represents a small contact surface with increased thermal resistance compared to the overall daughterboard surface. The resistive connection is further declined by the MyriadRF PCB that represents a thermal resistance between the LMS6002D chip and the four mounting holes.

The first-generation SDR design uses five important countermeasures to improve the weak thermal design of the MyriadRF so that the SDR and SEPP temperature is kept within the operational temperature limit:

- The gap between the MyriadRF daughterboard and the motherboard is filled with a thermal gap filler material. The gap filler maximizes the contact surface and improves the heat flow to the motherboard.
- Unused mezzanine connector pins between the MyriadRF and the motherboard are connected with multiple vias to the motherboard ground planes to achieve a low resistive thermal connection between the MyriadRF and the motherboard ground planes.
- All critical components of the motherboard are soldered onto thermal ground pads.

- The motherboard uses six ground planes and so-called PCB edge contacts to increase thermal conductivity between the PCB components and the PCB edges that have direct contact with an Aluminium housing.
- An Aluminium housing connects the PCB edge contacts with the satellite structure. The satellite structure uses a body-mounted radiator panel that emits the energy into space.

Figure 7 shows the first-generation OPS-SAT SDR motherboard and MyriadRF daughter board and the thermal gap filler material between the MyriadRF board and the motherboard. The golden edges of the motherboard represent the aforementioned thermal edge contacts.

The countermeasures improve the thermal design significantly with the result that the temperature difference between the LMS6002D package and the motherboard decreases from about 15 °C to only 2-3 °C, as shown in our laboratory tests. Nevertheless, the countermeasures did not entirely cancel the disadvantages of the MyriadRF daughter board and resulted in an increased design and test effort with additional costs. This fact is a reason to develop a new improved SDR design with a better thermal performance.

Fig. 7. The first-generation SDR motherboard with the MyriadRF daughter board. A thermal gap filler material is inserted between the two PCBs and special edge contacts are used to improve the thermal design.

F. Mechanical Design

The mechanical design of the second-generation SDR is essential for its overall performance and reliability, considering the following three aspects:

- Mechanical vibration and shock events during launch and satellite separation can heavily damage satellite components and can lead to the total loss of a satellite. An appropriate mechanical design protects the device from damage caused by external mechanical forces like vibrations and uses mechanical structures that do not resonate and break during launch.
- The mechanical design can have a positive effect on the radiation shielding of the various electrical components. Especially low energetic particles can be blocked with metal plates and housing structures.
- A satellite system is often limited in its total mass so that the mechanical design should support lightweight mechanical structures whenever possible to reduce the overall mass of the system.

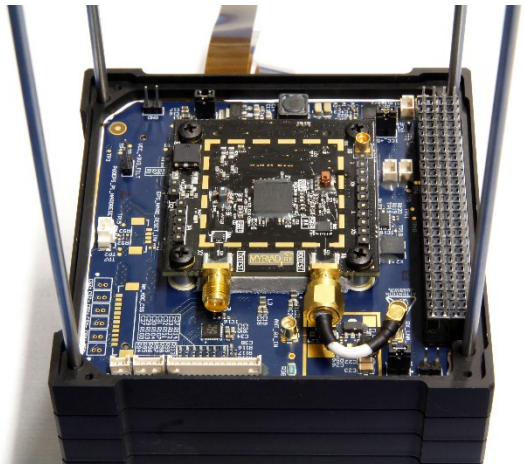


Fig. 8. The first-generation SDR motherboard with the MyriadRF daughterboard assembled into the flight housing frame.

Figure 8 shows the final assembly of the first-generation SDR inside the aluminium housing, with removed top cover plate. The housing frame improves the thermal energy transfer to the cold satellite structure and protects the components against low energy particles.

In addition to general design considerations, it is beneficial to perform a mechanical analysis, based on Computer Added Design (CAD) models, with the goal to understand the behaviour and reliability of the PCB in combination with the housing before manufacturing of the first SDR prototype. It is also required to perform vibration tests in accordance to the ESA/ECSS standards [8] with an assembled SDR prototype to verify the real behaviour of the SDR hardware and to identify any potential mechanical resonance that could damage the board during launch.

The MyriadRF is not designed for space and not manufactured as a highly reliable component, so that every single PCB has to be extensively tested to verify if it can survive the harsh space environment. We found out that especially the visual inspection and functional testing was very critical. Several potentially weak soldering joints and some mid-chip solder balls could be identified. Hence, every single PCB had to be reworked and cleaned to minimize the risk of solder joint defects during launch and electrical shorts between the components due to remaining solder balls. The vibration test results of the first-generation SDR revealed two natural resonance frequencies that are caused by the geometry of the MyriadRF PCB and the high mass in the centre of the motherboard PCB. The functional test confirmed that the PCB components were not destroyed or damaged during vibration testing. To reduce the risk of damages for the second-generation SDR, a PCB design with a more homogeneous mass distribution is going to be used.

III. The Second-Generation SDR for PRETTY

A. Overview

The experiences made during the hardware and software design, functional verification and device characterisation allows us to develop a new second-generation SDR for the PRETTY passive reflectometer and other future applications in space.

The second-generation SDR design for PRETTY implements the following list of improvements:

- The possibility to receiver very weak RF signal with a signal power of about -100 dBm.
- The functionality to perform full-duplex bidirectional communication between space and ground.
- The extension of the covered RX and TX frequency range to Global Position System (GPS) L1 band at 1.575 GHz [9] and S-Band at 2-4 GHz.
- An improved thermal design that increases the heat emission.
- An improved PCB design with a homogeneous mass distribution and reduced natural resonances.
- A clean and straightforward software design that allows the easy realization of the PRETTY passive reflectometer experiment and other signal processing features for bidirectional communication with the ground.
- The possibility to extend the SDR RF frontend with additional PCBs that enable the use of higher frequency bands.

In contrast to the first-generation SDR, all components of the second-generation SDR are selected with respect to their performance, their usability for space, the verification status and by taking into account the availability of design resources and documentation.

The second-generation SDR allows the reception of ultra-low power RF signals, which is required by the passive reflectometry, but can also be used as fully functional communication system for high-speed bidirectional communication.

It has to be emphasized that PRETTY does use a single RX channel only for the passive reflectometer while the other RX and both TX channels are not used. While it would be possible for several SEPP applications to access the SDR, it is dedicated to the reflectometer experiment. The PRETTY system requirements do not allow any signal transmission in parallel that could cause interferences or increased noise levels during the experiment. Nevertheless, the system can be used in contingency situations or as a backup system in case of problems with the nominal S-band transceiver. The PRETTY mission is an opportunity to test the reliability and performance of the second-generation SDR in space so that it achieves flight heritage for other missions.

B. The AD9361 RF Frontend

As core component, the Analog Devices AD9361 RF frontend chip is used for the second-generation SDR because we think that Analog Devices has an adequate quality assurance and product verification strategy to guarantee a good product quality. The company provides many design resources and documentation for the chip and software reference designs that support the work of the developers [10].

The AD9361 chip has flight heritage and was radiation tested [11] so that it is a perfect candidate for the second-generation SDR design.

Figure 9 shows the internal architecture of the AD9361 RF frontend chip and gives an impression on the advanced design complexity compared with the first-generation SDR. The chip contains two RX and two TX channels with quadrature de-modulators with a frequency range of 0.3 to 6 GHz. All RX and TX channels allow full-duplex communication with a channel bandwidth of up to 56 MHz.

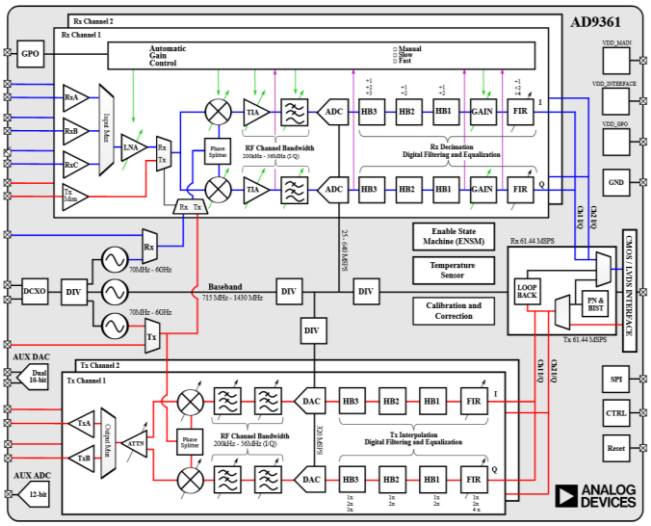


Fig. 9. The block diagram of the AD9361 RF frontend chip. The chip uses two pairs of independent RX demodulator and TX modulator channels and covers a frequency range of 0.3 to 6 GHz.

Configurable analogue and digital filters improve the signal conditioning and give high flexibility regarding the selection of the modulation scheme. A single clock source is used as reference for the internal RX and TX PLLs and the signal sampling, so that it is not required to generate and synchronize the different clock sources with an external clock generation, with the additional benefit of a simpler PCB design. Both RX channels are equipped with a configurable AGC functionality that helps to compensate atmospheric attenuation effects. The AD9361 uses built-in calibration methods with status and error checking, which is a significant improvement compared to the LMS6002D frontend.

C. System Architecture

Figure 10 shows the SDR RF frontend PCB architecture consisting of a high-frequency analogue portion and a low-frequency digital portion.

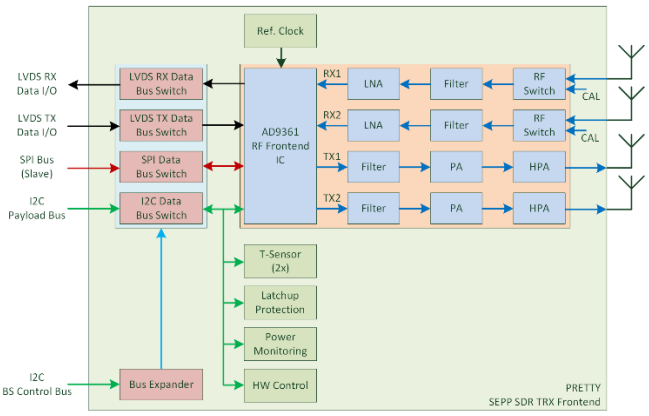


Fig. 10. The archtecture of the second generation SDR RF frontend board consists of two RX and two TX channels, some peripheral components and data bus switches for the baseband data I/O, configuration and monitoring interfaces.

The blue blocks in Figure 10 represent the high-frequency portion, consisting of an AD9361 RF frontend chip as the core component with two TX channels and two RX channels. Each

TX channel uses an RF filter to restrict the modulated TX signal spectrum to the allowed frequency band and two power amplifiers in series, to increase the TX signal power to about 1W per channel. Each RX channel uses an RF filter and an LNA to filter and amplify the received signals before demodulation and signal sampling. The LNA is primarily required to amplify very weak signals that lie below the AD9361 sensitivity threshold. The high-frequency portion contains everything that is required for the passive reflectometer system but allows applications to implement a full-duplex bidirectional communication on frequency bands like UHF, L-Band or S-Band.

The green and red blocks in Figure 10 represent the low-frequency portion, consisting of control and monitoring components, latch-up protection, power monitoring, two temperature sensors and a 40 MHz reference clock generator. The red blocks show the various data bus switches that allow the separation of the data interfaces from the satellite bus, to prevent failure propagation and bus locks that could be caused by an SDR defect. The data bus switches were introduced in the first-generation SDR and are reused since they are a proven method to improve the satellite fail-safety.

It has to be emphasized that the SEPP motherboard is used again for baseband processing, control and monitoring of the SDR RF frontend board during all experiments. The passive reflectometer software for PRETTY is implemented in the FPGA portion of SEPP while the signal reception, demodulation and sampling are performed by the SDR RF frontend. The SDR RF frontend board and the SEPP motherboard act as a combined SDR device that is used for the PRETTY passive reflectometer experiment.

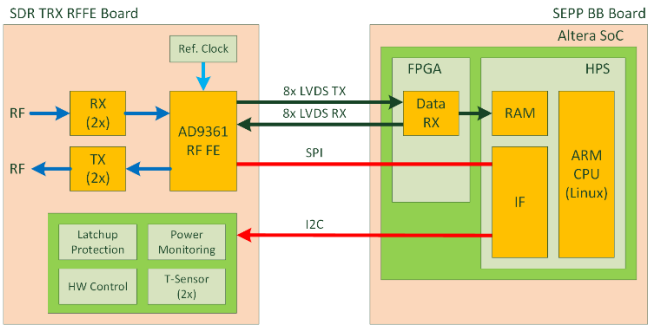


Fig. 11. The connections between the SDR RF frontend board and the SEPP motherboard. Eight LVDS pairs are used for the TX and the RX baseband data. An SPI master on SEPP controls and monitors the AD9361 chip. An I2C interface allows general RF frontend board control and monitoring.

Figure 11 shows the connections between the SDR RF frontend board and the SEPP motherboard. An 8-channel Low Voltage Differential Signalling (LVDS) interface allows high data rates up to 3.125 Gbps and improves the reliability of the baseband data exchange between the SDR and the SEPP due its differential signalling concept. LVDS produces less electromagnetic interference (EMI) and reduces the risk for bit failures due to EMI caused by other devices.

D. Software Design

The software design of the second-generation SDR differs entirely from the first-generation SDR because Analog Devices provides good software design support and software reference designs for FPGAs and Linux OS.

The SEPP API and high-level SDR software uses an existing AD9361 Linux OS driver from Analog Devices instead of hundreds of self-made low-level commands. The Linux driver interacts with the FPGA firmware and provides access to the baseband processing and Direct Memory Access (DMA) controllers inside the FPGA. Due to the custom design of the SDR RF frontend, some parts of the driver have to be reworked and updated to enable all features of the SDR. This process was simplified because of the availability of a reference implementation.

Standard interfaces like Industrial Input/Output (IIO) [12] allow easier extension and integration of additional FPGA and Linux OS components. The use of existing and industrial software concepts reduces the development time, guarantees a better design support and makes the second-generation SDR compatible with third-party software products.

E. Thermal Design

The thermal design of the second-generation SDR mitigates the disadvantages of the first-generation SDR design by avoiding the use of a daughter board for the RF frontend. Instead, all components with a high power dissipation are soldered directly onto the SDR motherboard to implement the best possible thermal connection to the internal PCB ground planes. Whenever possible, component packages with integrated thermal pads are selected to reduce the thermal resistance. All component ground pads and pins are connected with multiple ground vias that maximize the heat flow from the packages to the PCB ground planes.

The PCB layer-stack was increased from 10 layers to 14 layers with six ground planes to guarantee a low thermal resistance and a high thermal energy flow to the PCB edge contacts. A side effect of the higher layer count is an increased PCB thickness that yields to wider edge contacts with a larger contact surface and better heat transfer to the Aluminium housing. Due to the good performance of the OPS-SAT mechanical design, we decided to use the same housing concept for the second generation SDR. The housing implements again the thermal connection to the satellite structure and a body-mounted panel that radiates the thermal energy into the cold space.

During OPS-SAT thermal cycling and environmental testing, we observed that the separation of the SDR RF frontend PCB and the SEPP PCB leads a good distribution of the thermal energy and a lower overall temperature. The lower overall temperature is mainly caused by the two housing frames that have a much larger contact surface to the structure than a single PCB would have. Since both systems use the same housing, we expect the same effect for the PRETTY SDR hardware.

F. Mechanical Design

Figure 12 shows a 3D model of the PCB design. The second-generation mechanical design is simplified due to the use of a single motherboard without an additional daughter board.

The larger number of layers increases the overall thickness of the PCB and makes the PCB stiffer, resulting in a shift in geometry and mass-based resonances towards a higher resonance frequency. The resonance does not matter anymore if the resonance frequency exceeds the maximum frequency of the vibration spectrum that is present during the launch.

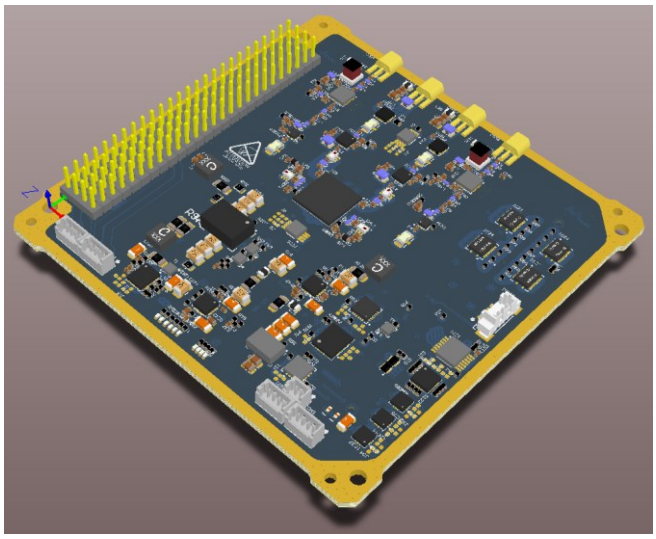


Fig. 12. The 3D model of the second-generation SDR RF frontend board design. The AD9361 RF frontend chip is located near the PCB centre close to the two RX and two TX channels.

Lower resonance amplitudes are expected compared the first-generation SDR due to the homogenization of the mass distribution and reduction of the mass in the centre of the PCB. The PCB is again mounted into a housing frame that holds the PCB in its position. The housing frame is fully compliant with the CubeSat Design Specification [13] so that it perfectly fits into the mechanical structure of the satellite.

All critical components are additionally spot bonded with an epoxide adhesive to reduce the risk for joint failures due to mechanical stress. The adhesive generates a stable connection between the component packages and the PCB surface that reduces the force onto the solder joints and the package pins. The resulting larger contact area has might have a positive impact onto the thermal design.

IV. Conclusion and Future Work

The design and architecture of the second-generation SDR for PRETTY bases on the experiences made during development and testing of the first-generation SDR for OPS-SAT. The second-generation SDR design uses an AD9361 RF frontend chip with two RX and two TX channels instead of a single receive channel. The second-generation SDR is ideally suited for the implementation of the PRETTY passive reflectometer because the design enables the reception of ultra-low power RF signals down to -100 dBm. The second-generation SDR design allows also full-duplex bidirectional data communication between the satellite and ground on frequency bands between 0.3 and 6 GHz with a two times higher modulation bandwidth of about 56 MHz compared to the first-generation SDR.

The use of metal housing frames follows a flexible modular design approach based on stackable PCBs of equal size. The PRETTY SDR configuration uses a single SDR motherboard that is stacked on top of two redundant SEPP motherboards. This configuration can be further extended in the future by adding RF extension boards for other purposes like the use of higher frequency bands.

The mechanical and thermal design is improved due to the absence of an RF frontend daughter board.

The engineering models of the second-generation SDR are currently under production and their functionality and reliability are going to be tested in the next weeks. Especially the characterization, vibration and thermal tests are going to be very important because they are the primary source of performance and reliability data. In parallel to these activities, the software and experimental setup is going to be implemented and tested by using the SDR engineering models.

All these verification results are the foundation for the final assembly of the SDR flight model and integration of the PRETTY satellite system, before it is launched in 2022.

References

[1] M. Martín-Neira, “A passive reflectometry and interferometry system (PARIS): application to ocean altimetry”, ESA J 17:331-355, 1993.

[2] S. T. Lowe, J. L. LaBrecque, C. Zuffada, L. J. Romans, L. E. Young and G. A. Hajj, “First Spaceborne observation of an Earth-reflected GPS signal”, Radio Sci 37(1):7-1-7-28, 2002.

[3] M. Martin-Neira, M. Caparrini, J. Font-Rossello, S. Lannelongue and C. S. Vallmitjana, "The PARIS concept: an experimental demonstration of sea surface altimetry using GPS reflected signals," in IEEE Transactions on Geoscience and Remote Sensing, vol. 39, no. 1, pp. 142-150, Jan. 2001.

[4] D. Evans, O. Koudelka, L. Alminde and K. Schilling, “The ESA OPS-SAT CubeSat Mission,” Majorca Island, Spain, 2014.

[5] MyriadRF.org, “The MyraidRF LMS6002D Reference Development Kit”, 08-Dec-2015. [Online]. Available: https://myriadrf.org/projects/component/rdk/. [Accessed: 02-Feb-2020].

[6] MyraidRF.org, “Lime Microsystem LMS6002D Product Datasheet”, 2020. [Online]. Available: https://wiki.myriadrf.org/LimeMicro:LMS6002D_Datasheet. [Accessed: 17-Apr-2020].

[7] James C. Maxwell, “Theory of Heat”, Greenwood Press, Westport, USA, third edition, 1970.

[8] ESA ESTEC, “TEC-SY/128/2013/SPD/RW - Tailored ECSS Engineering Standards for In-Orbit Demonstration CubeSat Projects”, 24-Nov-2016. [Online]. Avalable: https://blogs.esa.int/philab/files/2019/11/AD-01_IOD_CubeSat_ECSS_Eng_Tailoring_Iss1_Rev3.pdf. [Accessed: 20-Apr-2020].

[9] B. W. Parkinson, P. Enge, P. Axelrad and J. J. Spilker Jr., “Global Positioning System: Theory and Applications, Volume II”, American Institute of Aeronautics and Astronautics, 1996.

[10] Analog Devices, “AD9361 HDL reference designs”, 2020. [Online]. Avalable: https://wiki.analog.com/resources/eval/user-guides/ad-fmcomms2-ebz/reference_hdl#ad9361_hdl_reference_designs. [Accessed: 17.Jun.2020].

[11] J. Budroweit, M. Jaksch, R. G. Alía, A. Coronetti and A. Kölpin, Heavy Ion Induces Single Event Effects Characterzation on an RF-Agile Transceiver for Flexible Multi-Band Radio Systems in NewSpace Avionics”, Aerospace 2020, 7(2), 14, 2020.

[12] Analog Devices, “AD9361 HDL reference designs”, 2020. [Online]. Avalable: https://wiki.analog.com/software/linux/docs/iio/iio. [Accessed: 17.Jun.2020].

[13] California Polytechnic State University, “CubeSat Design Specification”, Rev. 13, 20. February 2014. [Online]. Available: https:// www.cubesat.org/s/cds_rev13_final2.pdf. [Accessed: 20-Feb-2020]

A GPS Patch Antenna Array for the ESA PRETTY Nanosatellite Mission

Reinhard Zeif
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
reinhard.zeif@tugraz.at

Andreas Hörmer
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
hoermer@tugraz.at

Manuel Kubicka
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
manuel.kubicka@tugraz.at

Maximilian Henkel
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
henkel@tugraz.at

Otto Koudelka
Institute of Communication Networks and Satellite Communications
Graz University of Technology
Graz, Austria
koudelka@tugraz.at

***Abstract*—The PRETTY mission is a satellite mission of the European Space Agency (ESA) with the goal to demonstrate the concept of passive reflectometry with a small and cost-efficient 3U Nanosatellite. Passive reflectometry allows the characterisation of the Earth surface properties and height, by correlating direct and reflected Global Positioning System (GPS) signals. The PRETTY mission focuses on the characterisation of the surface height. The system is realized as a Software Defined Radio (SDR) with a high-gain GPS patch antenna array that receives the very weak reflected signals from Earth's surface. The GPS patch antenna array resides on the outer surface of the satellite with direct contact to the harsh space environment. The application in space is a key factor for many requirements and design restrictions for the PRETTY GPS antenna array. This paper focuses on the various design constraints and requirements for the antenna array and describes the relations with the antenna design parameters. The paper describes the results obtained from a first single patch simulation and discusses the final antenna array architecture, simulation results and PCB design.**

Keywords—PRETTY, Passive Reflectometry, Antenna Array, Nanosatellite, Software Defined Radio, GPS

I. Introduction

The ESA PRETTY Nanosatellite mission uses a state-of-the-art Software Defined Radio (SDR) with a high-gain patch antenna array to demonstrate the concept of passive reflectometry [1]. The passive reflectometry was successfully tested in space in 2002 [2] and is realized now on a much smaller and more cost-efficient Nanosatellite, to show that this technology can be realized with low-cost components and reduced size and mass. PRETTY will be the first ESA mission that flies a passive reflectometer with a patch antenna array on a 3U Nanosatellite.

A passive reflectometer measures the characteristics of a surface by receiving and correlating two electromagnetic signals coming from a single external source. The first signal is received directly from the source, whereas the second signal is the indirect reflection from the surface.

Passive reflectometer systems require less power because they do not actively generate and transmit signals. Instead, electromagnetic signals from other external sources, e.g. other satellites, are reused. The passive design yields a smaller satellite size with reduced weight, a more cost-efficient mission, lower design complexity and shorter development time.

The PRETTY mission uses a single satellite on a Low Earth Orbit (LEO) to receive signals reflected by the Earth surface. The general concept of passive reflectometry is shown in Figure 1.

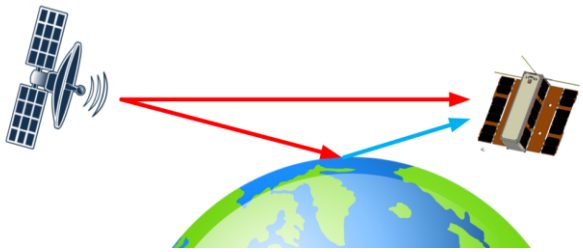


Fig. 1. The concept of passive reflectometry using a direct and reflected signal from an external signal source.

The PRETTY satellite uses Global Positioning System (GPS) L1 signals for the passive reflectometer with the goal to characterize the Earth surface height. GPS L1 signals are a good choice due to their well-known signal characteristics, relatively low carrier frequency of about 1.575 GHz and a large number of satellites, which result in good ground coverage [3]. The ground coverage is important to be able to receive a reflected signal all over the world so that the signals can be captured at any time independent of the orbit.

The PRETTY passive reflectometer is made of a SDR with an Analog Devices AD9361 radio frequency (RF) frontend and an Altera Cyclone V Field Programmable Gate Array (FPGA). The SDR has a very powerful amplifier stage with up to 86dB gain but even with this system, it is challenging to receive the direct, non-reflected GPS L1 signals and the very weak reflections. The signal power of the reflected GPS signal can be as low as -140dBm, so that the signal reception is very challenging, especially when the signal power is below the noise floor. The reception of the reflected signal requires a powerful receiver hardware and an antenna with a high gain. This fact makes the antenna to a core element of the passive reflectometer.

The PRETTY passive reflectometer uses a body-mounted patch antenna array since this is a reliable and space-proven

antenna concept that allows the realization of the required antenna gain.

This paper describes in chapter II the relevant satellite system requirements and constraints. Chapter III describes the various requirements and constraints and their relation to the antenna design parameters. Chapter IV discusses the simulation results of a first single antenna patch prototype. Chapter V discusses the final antenna array architecture with its components. Chapter VI discusses the simulation results of the antenna array architecture. Chapter VII describes the final PCB design used for the antenna array.

II. System Design

The GPS antenna array is a subsystem of the PRETTY satellite and has to be designed with respect to the selected satellite platform, the system requirements and the given design constraints.

The following sections describe the most important system requirements and design constraints.

A. System Requirements

The following list summarizes the most important system requirements for the PRETTY SDR representing the passive reflectometer:

- Reception of GNNS (GPS) L1 signals at 1.575GHz
- > 105dB SDR receiver gain, resulting in a total antenna gain of at least 20dB
- < 5dB in-band gain variation (+/- 10MHz)
- Bandwidth of +/- 10MHz at L1
- Total Noise Figure < 4 dB

The system requirements were selected in the ESA project phase A/B in accordance to the mission requirements and scientific goals.

The antenna has to be seen as critical element of the SDR frontend because it is the first element in the receive chain. The antenna design has a large impact onto the overall performance of the SDR and plays an important role for the fulfilment of the system requirements.

B. System Design Constraints

The satellite system design and the harsh space environment strongly influence and constrain the antenna array design. It is beneficial to identify the various design constraints to be able to define some design rules. The design rules act later as general guidelines for the actual antenna design process.

The following eight constraints are worth to mention:

1) Mounting position of the antenna: The most obvious limitation concerns the selection of the satellite side on which the antenna is mounted. The antenna has to be located on an electrically grounded outer surface of the satellite, so that the structure and the mechanical parts of the satellite do not adversely affect the antenna. Large mechanical parts shall not interfere with the antenna or affect its characteristics.

During project phase A and B, many system design iterations were made to define the nominal attitude of the satellite with the goal to point the satellite to the sun to charge the batteries, but also to allow communication with a ground station and the execution of the passive reflectometry experiments. All these aspect taken into account, it was decided to use a body-mounted panel on the backside of the satellite for the antenna array. The PRETTY 3U outer surface with the body-panels and solar wings is shown in Figure 2.

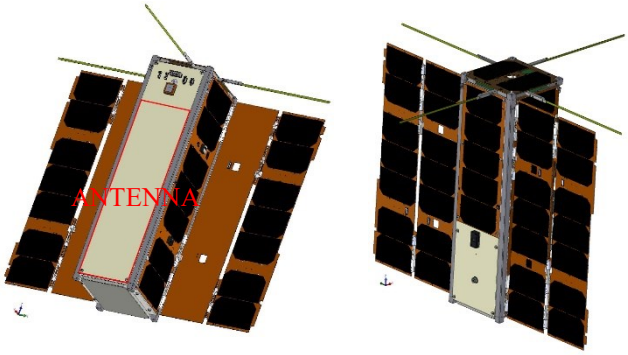


Fig. 2. The PRETTY 3U satellite backside with its two half-populated solar panel wings (left) and the front side with fully populated solar wings (right). The antenna for the passive reflectometer is mounted onto the backside (left).

Placing the antenna array onto the backside of the satellite has the positive effect that the antenna is not illuminated and not heated up by the sunlight. The surface temperature and the thermal noise level are thereby at a minimum.

2) Mechanical Dimensions: The size and shape of the satellite structure restrict the antenna size. The antenna can only be mounted onto a body-mounted satellite panel. The antenna must not exceed the given size and height restriction so that the satellite is compliant with the deployment pod. For PRETTY, the maximum available space for the antenna is 80 x 200 x 8mm. These constraints restrict the number of the array patches, which in turn affects the achievable total gain. Additionally, the RF bandwidth is typically limited by the allowed patch antenna height.

3) Mass: The allowed total satellite mass restricts the mass of the antenna. The mass is very critical since most deployments pods and mechanisms are qualified for a satellite mass of 4kg [4]. Considering the mass budget for the rest of the satellite, the mass of the antenna is limited to 400g.

4) Center of Gravity (CoG) and Moment of Inertia (MoI): The CoG of the satellite has to be close to the geometric centre. A large and heavy antenna shifts the CoG away from the geometric centre and changes the MoI. This puts additional demands on the satellite attitude control system.

5) Reliability of Mechanisms: It would be possible to increase the antenna array size and gain by using a deployable antenna. Unfortunately, the deployment mechanisms are complex and error-prone. Due to the importance of the antenna for the success of the PRETTY mission, it was decided not to accept the additional risk.

6) Component Selection: Component selection is very critical for space missions. Not all components are designed for the harsh space environment, with the consequence that there is a potential risk for failure. The failure risk can be reduced by using qualified components, testing and screening [5]. Hence, a compromise between performance, cost, reliability and availability of components has to be found. Due to the low-cost aspect of the PRETTY mission, only Commercial Of-The-Shelf (COTS) components are used.

7) Thermal Design: The satellite thermal design is a demanding task during system design and essential to keep

the temperature inside the satellite in the allowed operational range. Figure 2 shows the front side of the satellite that is fully populated with solar wings. The solar wings are nominally operated in the so-called sun-pointing mode so that the solar wings are fully illuminated by the sun. The sun pointing maximizes the amount of generated electrical power but leads to an increased satellite surface temperature. Additionally, the satellite is heated up internally by the thermal loss of the electronic components. Hence, the remaining body-mounted panels on the shadowed satellite surfaces are required to dissipate the thermal energy. Since the antenna array is nominally one of the shadowed body-mounted panels, this panel has to contribute to the overall heat dissipation towards cold space. A key factor for this is the choice of proper materials and conformal coatings to achieve a high thermal dissipation infrared energy and low absorptance of incoming solar energy.

8) Radiation Design: The space environment with its various radiation sources can damage integrated circuits and affect material properties. The radiation intensity varies with satellite orbit altitude and as such, the chosen orbit is an essential factor for the mission and component lifetime. Shielding materials, like aluminium plates or coatings, are used to protect the components inside the satellite. The PRETTY mission foresees an orbit height in the range of 400 to 600 km, with a relatively low radiation total dose of less than 20 krad per year. Due to the low radiation rate, the short mission duration and the low-cost aspect of the PRETTY mission, only COTS components are used instead of the more expensive and more reliable space-grade or radiation hardened components.

III. Relations Between Requirements And Design Parameters

A. Overview

Before development of a first antenna prototype, the given requirements were sorted by their criticality and related to the system constraints, to identify the most critical requirements and design parameters. Some of the design parameters interdepend on each other, which reduces the degree of freedom for the design.

The following sections list the critical requirements in descending order of their criticality and provide a list of design parameters and their reciprocal relations.

B. Criticality of Requirements

The following critical requirements have a direct influence on the performance and the proper operation of the passive reflectometer, listed in descending order of criticality:

1) Antenna gain: A high total gain is the basis for the reception of the very weak reflected signals.

2) RF bandwidth: The antenna RF bandwidth has to be large enough to be able to receive the direct and the reflected GPS signals.

3) Noise Figure: Additional noise reduces the S/N ratio and makes signal processing more difficult. Hence, it was decided to achieve an antenna system Noise Figure (NF) of less than 3 dB.

4) Directivity: The antenna shall have a directed characteristic to allow signal reception from a selected fraction of Earth's surface. The final decision was made in favour of a Microstrip patch antennas array [6], since this technology provides the best match for the above-mentioned criteria.

C. Design Parameters

Especially the directivity and gain requirement, which luckily have a positive contribution to each other, were the key drivers for selecting a patch array as antenna architecture. Some design factors underpinned this decision:

1) Size versus wavelength: The GPS L1 signal wavelength $\lambda_{L1} = c \cdot f_{L1} = 19.05\,cm$ (in vacuum) is comparably large to the satellite size, if $\lambda/2$ and $\lambda/4$ antennas are used. Hence, the use of antenna dipoles or helix antennas, with a perpendicular orientation to a ground plane, is no option. Printed antennas, like Stripline patch antennas, have a much lower height but are large, if substrates with a low relative permittivity ε_r are used.

2) Size versus gain: An antenna array or printed Microstrip antennas with a high gain are required for PRETTY, but some of these designs do not comply with the size constraints.

3) Bandwidth versus height: The bandwidth of a patch antenna is influenced by various factors like the height and the relative permittivity ε_r of the substrate [7].

4) Radiation shielding versus antenna characteristics: Shielding is the simplest but not very effective way of radiation protection since it requires much mass and cannot be applied everywhere. For example, coatings or foils with metals are unsuitable for antenna patches because this would affect their characteristics. Nevertheless, the body-mounted panel itself, on which the antenna is mounted, contributes to the radiation shielding of the internal satellite components if it contains some shielding layer.

Because of 1) and 3), it was decided to use materials with a high relative permittivityε_r. The resulting reduction of the antenna patch dimensions allows the use of more patches for the antenna array, which contributes to a higher gain.

IV. Patch Antenna Simulation

After deciding to use a Microstrip patch antenna array, a simulation of a single patch was made, to verify, if the gain and bandwidth requirements can be fulfilled.

These early simulation was based on a self-made S-Band patch antenna design, manufactured in house for the OPS-SAT satellite mission [8]. For this purpose, the size of the antenna patch was tuned to the GPS L1 frequency band.

The patches consist of a 1mm thick copper ground plate and a 1mm thick copper patch. The space between the patch and the ground plate is filled with Rohacell material. As a next step, the antenna patches were combined with ideal combiners to estimate the resulting total gain of the array. The antenna characteristic is shown in Figure 3.

The simulation of an ideal antenna array confirmed the feasibility of the general concept but unveiled some weaknesses that have to be taken into account:

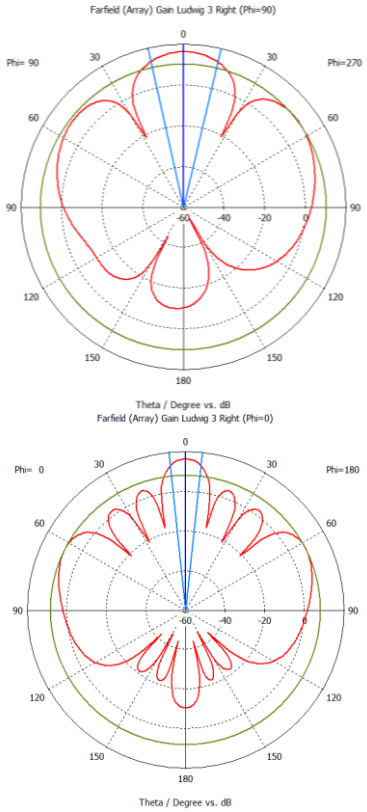


Fig. 3. Far field (f=1.580) simulation of the gain for an ideally combined and spaced antenna array consisting of eight patches. The ideal total antenna gain is about 16 dBi. This simulation confirms the feasibility of the concept and indicates that a high gain is achievable under realistic conditions.

1) Patch size and spacing: The usage of the Rohacell material ($\varepsilon_r = 1$) results in a large patch dimension. It turned out that the available surface area of about 80 x 200 mm is insufficient to place two patches next to each other while fulfilling the requirement that each patch is surrounded by a large ground plane. The simulations of these tightly arranged patches shows a reduced total gain and a change in the insertion loss, as described in [9]. Hence, it was decided to use smaller patches with a total ground plane size of two times the patch diameter.

2) Relative permittivity ε_r: The size of an antenna patch is proportional to the relative permittivity of the material between the patch and the ground plane. Consequently, an antenna with a large ε_r and smaller patch size is used to achieve a larger patch spacing.

3) Materials and manufacturing: The manufacturing process and available materials limit the design of a multi-layer PCB with integrated Microstrip patches and Stripline structures. The required antenna bandwidth results in a thick PCB with ten or more layers that are difficult and cost-intensive in production. Hence, we decided to use a PCB with fewer layers in combination with automotive-grade ceramic COTS antennas that are soldered onto the top layer. The initial idea of using self-made copper patches was withdrawn to reduce the risk of unforeseen effects caused by manufacturing tolerances.

Figure 4 shows the design that would result from the simulated antenna patches. Based on these first results, an improved antenna array architecture was defined in order to mitigate the identified weaknesses of the initial design.

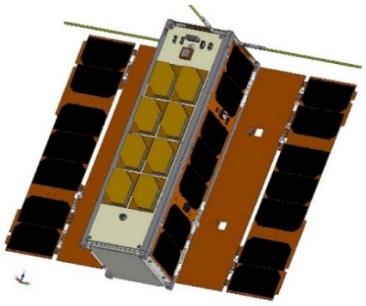


Fig. 4. The first draft of the antenna array consisted of eight patches. The patch dimensions were large and did barely fit onto the body panel resulting in a very low patch spacing.

The following section describes the improvements and the final architecture and design of the manufactured antenna array prototype.

V. Antenna Array Architecture

A. Overview

The antenna array architecture shown in Figure 5 was defined under consideration of the first simulation results.

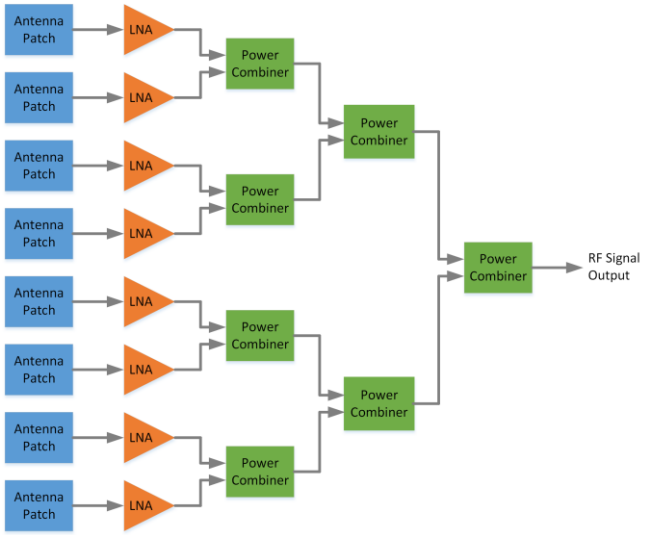


Fig. 5. The patch array architecture with eight patches, eight low noise amplifier and seven power combiners. The bias-T at the RF signal output for the LNA supply is not shown.

Eight GPS L1 patches represent the first element of the antenna architecture. The patches are mounted onto the outer surface of a body-mounted satellite panel, arranged in a 2x4 configuration, as shown in Figure 4. Low Noise Amplifiers (LNAs) are connected to each patch antenna feed point. The LNAs provide the additional gain to achieve the required total antenna gain. In an important final step, the patches are pairwise coupled via power combiners to a single 50Ω matched RF output signal [10].

It has to be emphasized that special care was taken on the concept and design of the satellite body panel. The goal of this process was to find a solution that allows the:

- Mounting of all required components, including the LNAs and peripheral components
- Usage of impedance-matched signal traces

- Realization of power combiners with small form factor but good separation and low insertion loss
- Achievement of the smallest possible overall form factor
- Consideration of the space environment

Due to the reliability, simplicity and comparable low manufacturing tolerances, it was decided to use a multi-layer Printed Circuit Board (PCB) for the body panel. More information on this body panel PCB is given in the following section.

B. *Body Panel PCB*

The body panel is a 10-layer Multi-Layer PCB made of Rogers RO4350B substrate with a thickness of 1.5mm.The PCB carries all the antenna patches and required SMD components like the LNAs and passive inductors or capacitors. The antenna patches are soldered onto the top layer. The remaining SMD components are soldered onto the bottom layer. Figure 6 shows the layer stack-up of the PCB and the used Striplines.

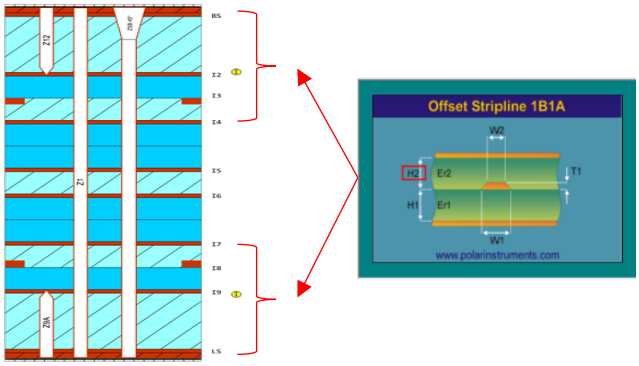


Fig. 6. PCB layer stack definition with eight signal and two constructive layers (left). The PCB uses a RO4350B substrate with a relative permittivity of about 3.5.

All layers are designed for 50Ω impedance-matched traces. The two Stripline layers 2 and 9 are designed for 50Ω and 70.7Ω impedances. Blind vias from top and bottom to the close by Stripline layers allow the stub-free signal routing. A power supply plane on layer 5 with a good capacitive coupling to a close ground plane provides the power for the LNAs.

The five power planes are required for the Striplines and the power supply. These planes have a high contribution to the thermal design because of their low thermal resistance. Additionally, they provide an additional 0.08mm copper for radiation shielding.

C. *GPS L1 Ceramic Patches*

Eight identical patch antennas with a low production tolerance are required to achieve a good array performance. A Taoglas DSGP.1575.25.4.A.02 surface-mount GPS L1 ceramic patch was selected. The patches have a small size of 25 x 25 x 4mm, due to the ceramic substrate ($\varepsilon_r \approx 40$). The peak gain is 4.3dB with a bandwidth of about 20MHz [11].

D. *Low Noise Amplifier*

The Minicircuits PMA2-33LN+ LNA was selected because of its tiny size of 2 x 2mm and ultra-low noise figure of 0.38dB. The LNA provides a gain of about 15dB [12].

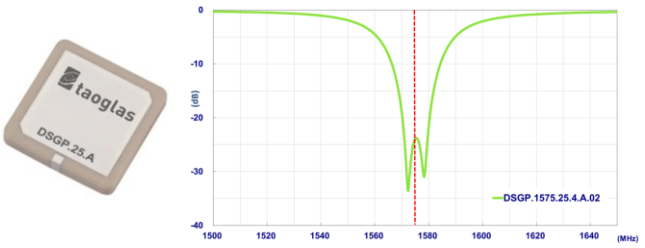


Fig. 7. Taoglas DSGP.1575.25.4.A.02 GPS L1 ceramic antenna patch (left). The antenna return loss shows a bandwidth of about 20MHz at -10dB (right) [11].

E. *Power Combiner*

The LNA output signals are pairwise combined with Wilkinson Stripline power combiner. These combiners provide adequate performance and have a small form factor.

VI. Simulation Results

A simulation with AWR Microwave Office was made to design the Stripline power combiners and to estimate the performance of the selected array architecture. The simulation was started with the design and evaluation of various Wilkinson power combiner Stripline layouts to find a variant with an insertion loss larger than -3.5dB and a small footprint of less than 30x30mm.

Next, the characteristics of the existing antenna signal paths were simulated, including the LNAs and Wilkinson combiners. This was necessary to verify if the requirements can be fulfilled and if the design can be realized as PCB. Additionally, the confidence level for the correctness of the Stripline dimensions and layer stack concept was improved.

The following sections describe the most relevant simulation settings and results.

A. *Low Noise Amplifier*

The LNA was approximated by using a linear amplifier that provides a gain of 14.5dB at 1.575GHz. The gain was reduced by 0.3dB to consider some possible loss margin.

B. *Wilkinson Power Combiner*

Figure 8 shows the selected Wilkinson power combiner layout. The combiner uses meander structures for the $\lambda/4$ traces to reduce the footprint. The meander spacing is more than factor 8 of the Stripline height to prevent crosstalk of the traces.

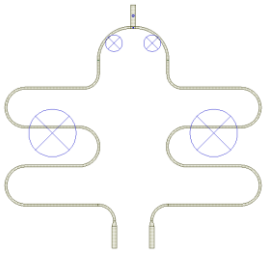


Fig. 8. Meander structure of the Stripline GPS L1 1.575GHz Wilkinson power combiner. The meander structure reduces the overall size. Input and output ports are matched to 50Ω.

Figure 9 shows the schematic used for the simulation of the Wilkinson combiner.

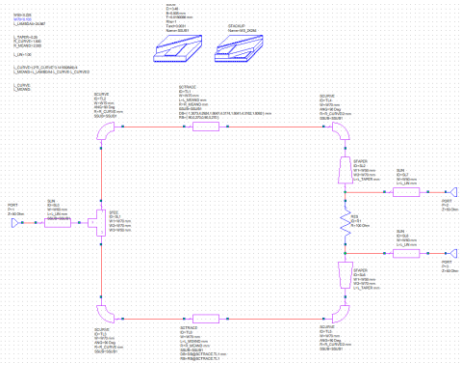


Fig. 9. Schematic used for the simulation of the Stripline GPS L1 1.575GHz Wilkinson power combiner. The simulation was performed with the NI AWR Design Environment 14 software.

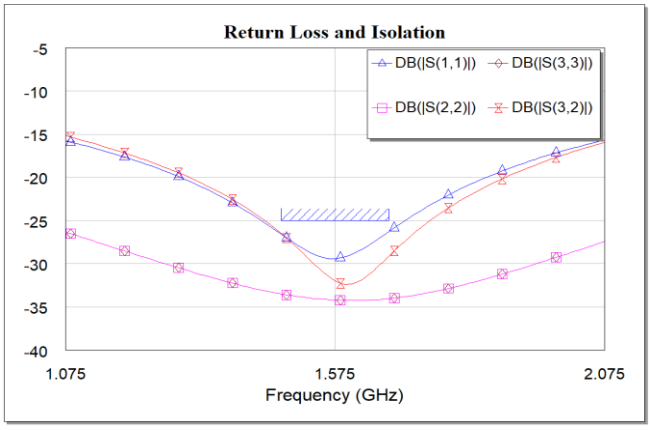


Fig. 10. The return loss of the Wilkinson power combiner is smaller than the design goal of -25dB for a bandwidth of 100MHz. The same applies to the isolation between the combiner ports 2 and 3.

Figure 10 shows the simulation results obtained with the NI AWR Design Environment 14 software. The simulated return loss and the isolation fulfil the design target of -25dB for a bandwidth of 100MHz. The simulated insertion loss is -3.2dB, equivalent to a combiner gain of 2.8dB.

C. Signal Path Gain

The eight RF signals from the patch antenna feed points are amplified by a LNA and combined to a single RF output signal, as shown in Figure 5. Hence, the signal path of each patch antenna consists of a LNA followed by three Wilkinson combiners. The gain and isolation of this signal path are shown in Figure 11.

The total signal gain is 23.1dB under the assumption that the RF input signals at each antenna feed have a power of 0dBm. It has to be emphasized that this gain does not take the antenna patch gain and the array factor into account.

As a conservative guess, an additional PCB loss of 0.6dB is assumed for the PCB, resulting in a signal path gain of 22.5dB.

Due to the combination of the eight patches, the antenna patch gain of 4.3 dB is doubled three times [8], which adds another 9dB. As a result, a total antenna gain of g_{total} = 22.5 + (4.3 + 9) = 35.8dB can be achieved.

The isolation of two different antenna inputs is between -33 and -45dB, depending on the number of combiners in the signal path.

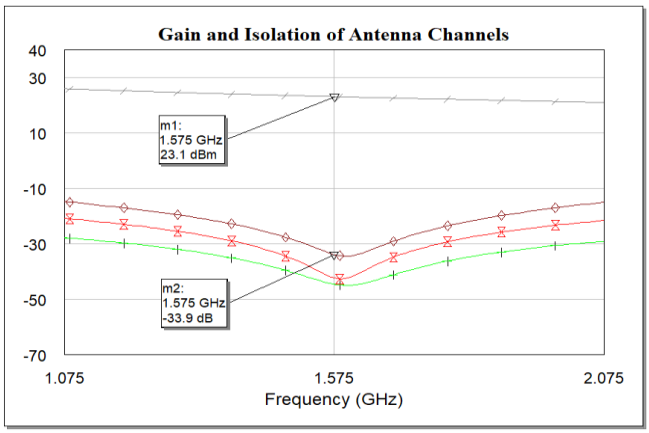


Fig. 11. Gain and isolation of the eight antenna channels, including the LNA and power combiners. The gain plot does show the signal path gain for antenna input signals of 0dBm. The isolation between two antenna inputs is between -33 and -45dB.

VII. The Antenna Array PCB

The antenna array multi-layer PCB uses a Rogers RO4350B substrate with 10 layers. Two of the 10 layers are constructive layers so that the PCB consists of eight layers. The 3D model of the PCB is shown in Figure 12.

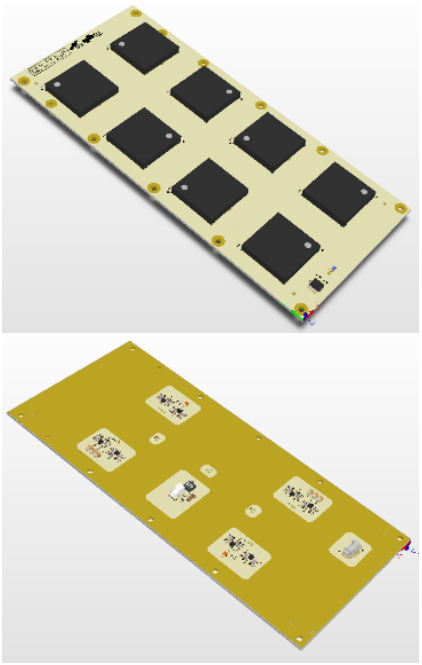


Fig. 12. The 3D model of the antenna array PCB. The top layer on the outer side of the satellite contains the eight antenna patches and some sensors (left). The bottom side contains the LNAs, RF antenna output connector and some peripheral components (right).

The PCB top side contains the eight antenna patches and builds the ground plane for the patches. Only a single impedance optimized via is used to connect the antenna feed point to the LNA inputs on the PCB bottom layer. The centre distance of two antenna patches is 43mm (= $0.225 \cdot \lambda_{L1}$ in vacuum) resulting in a ground plane of at least 40x40mm (= $2 * W_{PATCH} = \lambda_{PATCH\_L1}$ with ceramic substrate).

The LNA circuits are placed onto the bottom layer. The power combiners for the eight LNAs are realized on an inner Stripline layer. The combined RF output signal is routed to an

RF connector. A bias-T network is used to supply the LNAs via a coaxial RF cable.

Figure 13 gives an overview of the Wilkinson combiner structures that connect the LNA outputs to the RF output connector.

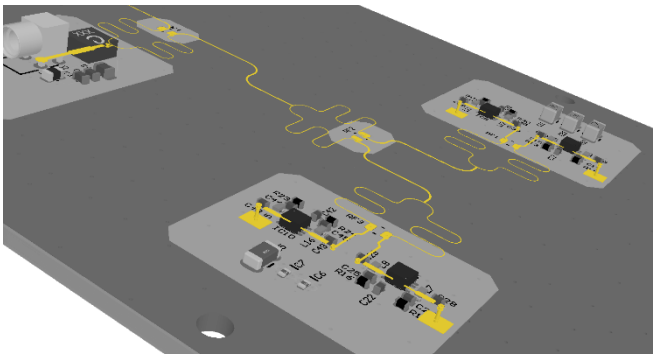


Fig. 13. The component islands consisting of two LNAs with peripheral components and the used RF structures with the Stripline Wilkinson power combiners.

Special attention should be given to the sophisticated layer stack and blind via concept that allows the routing of RF signals without stubs. The prevention of stubs in the signal chain reduces negative parasitic effects, contributes to a good impedance matching and improves the signal integrity. Figure 14 gives detailed insight into the stub-less RF signal routing between the top layer antenna feed, the bottom layer LNAs and the Wilkinson combiner on the inner Stripline layer.

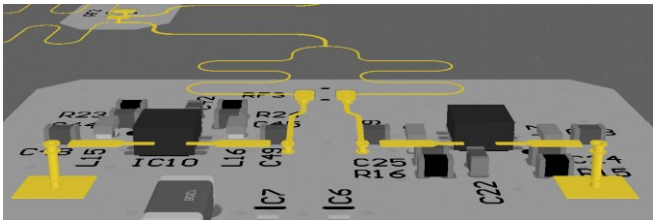


Fig. 14. A detailed view of the stub-less RF signal routing. Note that the PCB is rotated so that the top layer is on the bottom side. The antenna patch feeds (largely spared pads) are connected to the LNAs on the bottom layer and the Stripline combiner on the inner layer.

Another important point to mention is the fact that the bottom layer components interfere with the aluminium walls of the satellite structure, if no adequate countermeasure is implemented. Hence, a 2mm thick aluminium spacer plate is inserted between the PCB bottom layer and the satellite structure. This spacer plate has a positive impact on the thermal capacity, heat transfer and radiation protection.

VIII. Conclusion

We conclude from the obtained results that it is possible to design a patch antenna array with 36dB gain for a 3U Nanosatellite passive reflectometry mission under the given design and budgetary constraints. The implemented body-mounted RF PCB design with power combiners and ceramic patch antennas is a compact and powerful.

Furthermore, the simulations shows that there is some room to further increase the total gain by 6dB to 41.5dB if a LNA with a gain of 20.5dB is used.

IX. Future Work

The prototype of the antenna array PCB is currently being manufactured. This so-called engineering model is going to be verified and characterized in an anechoic chamber. Vacuum, thermal cycling, thermal shock, vibration and mechanical sustainability tests are made to qualify the system for flight. If required, further improvements are implemented and tested. Finally, a flight version is manufactured for the satellite.

References

[1] M. Martín-Neira, “A passive reflectometry and interferometry system (PARIS): application to ocean altimetry”, ESA J 17:331-355, 1993.

[2] S. T. Lowe, J. L. LaBrecque, C. Zuffada, L. J. Romans, L. E. Young and G. A. Hajj, “First Spaceborne observation of an Earth-reflected GPS signal”, Radio Sci 37(1):7-1-7-28, 2002.

[3] B. W. Parkinson, P. Enge, P. Axelrad and J. J. Spilker Jr., “Global Positioning System: Theory and Applications, Volume II”, American Institute of Aeronautics and Astronautics, 1996.

[4] California Polytechnic State University, “CubeSat Design Specification”, Rev. 13, 20. February 2014, URL: https://www.cubesat.org/s/cds_rev13_final2.pdf, 2014.

[5] NASA Engineering & Safety Center, The NESC 2014 Technical Update, Langley Research Center, NASA/TM-2014-218261, 2014.

[6] D.M. Pozar, “Microstrip Antennas”, IEEE Proceedings, vol. 80, pp. 79-91, Jan. 1992.

[7] R. B. Waterhouse, “Microstrip Patch Antennas: A Designer's Guide”, Kluwer Academic Publisher, Boston, 2003.

[8] D. Evans, O. Koudelka, L. Alminde and K. Schilling, “The ESA OPS-SAT CubeSat Mission,” Majorca Island, Spain, 2014.

[9] N. Minh Tuan, K. Byoungchul, C. Hosung and I. Park, “Effects of ground plane size on a square microstrip patch antenna designed on a low-permittivity substrate with an air gap”, 2010 International Workshop on Antenna Technology (iWAT), Lisbon, pp. 1-4, 2010.

[10] C. Bowick, J. Blyler, C. Ajluni, „RF Circuit Design”, Newnes Elsevier, 2011.

[11] Taoglas Limited, “DSGP.1575.25.4.A.02 Specification”, URL: http://cdn.taoglas.com/datasheets/DSGP.1575.25.4.A. 02.pdf, 2020.

[12] Minicircuits Inc., “PMA2-33LN+ Product Datasheet”, URL: https://www.minicircuits.com/pdfs/PMA2-33LN+.pdf, 2020.

DF Relayed OOK and PAM FSO Links with Turbulence and Time Jitter

P.J. Gripeos
Section of Electronic Physics and Systems, Department of Physics
National and Kapodistrian University of Athens, Athens, 15784, Greece
pgrypaios@phys.uoa.gr

H.E. Nistazakis
Section of Electronic Physics and Systems, Department of Physics
National and Kapodistrian University of Athens, Athens, 15784, Greece
enistaz@phys.uoa.gr

G.D. Roumelas
Section of Electronic Physics and Systems, Department of Physics
National and Kapodistrian University of Athens, Athens, 15784, Greece
groumelas@phys.uoa.gr

V. Christofilakis
Physics Department, Electronics - Telecommunications and Applications Laboratory, University of Ioannina, Ioannina, 45110, Greece
vachrist@uoi.gr

A.D. Tsigopoulos
Sector of Battle Systems, Naval Operations, Sea Studies, Navigation, Electronics and Telecommunications
Hellenic Naval Academy, Piraeus, 18539, Greece, atsigo@snd.edu.gr

G.S. Tombras
Section of Electronic Physics and Systems, Department of Physics
National and Kapodistrian University of Athens, Athens, 15784, Greece
gtombras@phys.uoa.gr

Abstract - In recent years, the increasing research and commercial interest for FSO communication systems has included them among the popular and effective communication technologies worldwide. Nevertheless, the main drawback of the terrestrial FSO links is related to the randomly time-varying atmospheric characteristics. In this work, the joint influence of time jitter effect and weak atmospheric turbulence, modeled with the gamma distribution, at the average BER performance of serially relayed Decode-and-Forward terrestrial FSO links, is investigated for two typical modulation schemes, i.e. OOK and PAM. The scope of this work is to extract accurate closed-form mathematical expressions for the system's performance estimation. Furthermore, the corresponding numerical results are presented for various typical FSO parameter values.

Keywords—Optical Wireless Communications; Terrestrial Free Space Optical Links; Decode-and-Forward Relays; Gamma Modeled Atmospheric Turbulence; Time Jitter; Modulation Format; Average BER Estimation

I. INTRODUCTION

In recent years, many state of the art applications are based on the optical wireless communications (OWC) – or Free-Space Optical (FSO) communication – systems, as one of the most popular wireless communication technologies worldwide. Besides the research interest of FSO, the commercial one is also high, due to the advantageous performance of these systems. In particular, FSO systems utilize the infrared and visible band of the electromagnetic spectrum, which is extremely wide, harmless to humans, and free of user license charge. These features enable FSO systems to achieve secure, reliable and high-rate communications with exceptionally low both installation and operational costs [1-4].

Nevertheless, FSO communications are highly affected by the current weather conditions of the propagation medium, the troposphere, which in combination with the scintillation effect may deteriorate the propagating signal up to the outage limit. Thus, the randomly time-varying characteristics of the FSO channels impose the use of a variety of statistical methods, in order to cope with the probabilistic uncertainty of the receiver's fluctuating signal power due to weather conditions and atmospheric turbulence effects, [5-12]. Depending on the prevailing turbulence conditions, a number of statistical models have been adopted in order to accurately support the expected results. Especially important, time jitter affects high data rate communications, because of the dicey short duration of the optical pulses used. Ignoring this effect, the receiver's detector is at high risk of incoming bit misjudgment, causing additional bit flip errors [13-23]. Taking all the above into consideration, it can be deduced that long, terrestrial FSO links are prohibitive, as the degradations are accumulated along the propagation path. This problem can be partially overcome by using a variety of techniques in order to control burst errors events, like decode-and-forward (DF) relay nodes between the initial transmitter and the final receiver, multi-channel diversity schemes or smart channel coding with forward error correction (FEC) [24, 25]. The last two configurations, i.e. diversity or FEC, can improve significantly the system's performance, but here are beyond the scope of our work. Thus, the first case is thoroughly investigated in this work by estimating the average bit error rate (ABER) for a fixed link length, between the initial transmitter until the final receiver, with a variable number of DF relays for two modulation schemes, i.e. on-off keying (OOK) or pulse amplitude modulation (PAM), using the Gamma distribution, which is appropriate enough for theoretical investigation of weak turbulence conditions, [26-30], and also the time jitter effect.

The remainder of this work is organized as follows: in section II, the model used is described and closed-form mathematical expressions of the average BER are derived while, in the next section, the corresponding numerical results are presented for typical FSO parameter values. Finally, the conclusions of this work are presented in section IV.

II. SYSTEM MODEL

A. Turbulence influence on each individual link

Supposing a stationary, ergodic and memoryless atmospheric channel exposed to additive white Gaussian

noise (AWGN), n, with zero mean value and variance σ_n^2, the receiving signal, r, is given as, [29-31]:

$$r = es_m I + n \tag{1}$$

where e denotes the effective photon-to-electron conversion ratio, s_m represents the modulated signal transmitted and I stands for the normalized irradiance of the channel, [28, 30]. In this work, the gamma distribution model is used and its probability density function (PDF) is given as, [26-28, 32]:

$$f_I(I) = \frac{c^c}{\Gamma(c)} I^{c-1} \exp(-cI) \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function and the coefficient $c = \left(a^{-1} + b^{-1} + ab^{-1}\right)^{-1}$, [26, 32], is related to the parameters of the link through the following expressions, [33-36]:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \left[\exp\left[0.49c_R^2\left(1+0.18p^2+0.56c_R^{12/5}\right)^{-7/6}\right]-1\right]^{-1} \\ \left[\exp\left[\dfrac{0.51c_R^2\left(1+0.69c_R^{12/5}\right)^{-5/6}}{\left(1+0.9p^2+0.62p^2c_R^{12/5}\right)^{5/6}}\right]-1\right]^{-1} \end{bmatrix} \tag{3}$$

where $c_R^2 = 0.5C_n^2 k^{7/6} R^{11/6}$ is the Rytov variance and $p = 0.5d\sqrt{k/R}$, C_n^2 is the structure index parameter varying from 10^{-13} $m^{-2/3}$ to 10^{-17} $m^{-2/3}$ for strong-to-weak turbulence effects, respectively, [35], d stands for the receiver's aperture diameter and $k=2\pi/\lambda$ is the optical wave number with λ and R being the operational wavelength and the length of the optical link, respectively. By performing a simple random variable (RV) transformation, the PDF of Eq. (2) can be equivalently written as:

$$f_\gamma(\gamma) = \frac{c^c}{2\Gamma(c)} \frac{\gamma^{\frac{c}{2}-1}}{\mu^{\frac{c}{2}}} \exp\left(-c\sqrt{\frac{\gamma}{\mu}}\right) \tag{4}$$

where $\gamma = e^2 I^2/\sigma_n^2$,[37, 39] and $\mu = e^2\left(E[I]\right)^2/\sigma_n^2$,[30, 39, 40] are the receiver's instantaneous and expected signal-to-noise ratio (SNR), respectively, while E[·] is the expectation operator, [41, 42]. Concerning the average BER estimation of the FSO link due to the gamma modeled turbulence effect, a generic formula, for any modulation scheme, is obtained as, [37, 38, 43, 45]:

$$BER_{Turb,s} = \frac{1}{2}\int_0^{+\infty} erfc\left(\sqrt{\gamma}/h_s\right) f_\gamma(\gamma) d\gamma \tag{5}$$

where erfc(·) stands for the complementary error function, [41], s indicates each specific modulation scheme which is used and the parameter h_s is taking the following values:

$$\begin{bmatrix} h_{RZ-OOK} \\ h_{NRZ-OOK} \\ h_{M-PAM} \end{bmatrix} = \begin{bmatrix} 2 \\ 2\sqrt{2} \\ 2\sqrt{2}(M-1)/\sqrt{\log_2 M} \end{bmatrix} \tag{6}$$

By substituting (4) into (5) and by transforming the exponential functions to the appropriate G-Meijer ones, [35, 45], the integral of (5) can be estimated analytically and the average BER is given as, [35, 36, 47]:

$$BER_{Turb,s} = A_s C(\mu) G_{2,3}^{2,2}\left(B_s(\mu) \left| \begin{matrix} \frac{2-c}{2}, \frac{1-c}{2} \\ 0, \frac{1}{2}, -\frac{c}{2} \end{matrix} \right. \right) \tag{7}$$

where $C(\mu) = c^c/\left(\pi\Gamma(c)\mu^{c/2}\right)$, with G[·] being the Meijer G function, [48], while the parameters A_s and $B_s(\mu)$ are taking the following values depending on the modulation scheme, [35, 46, 47]:

$$\begin{bmatrix} A_{RZ-OOK} \\ A_{NRZ-OOK} \\ A_{M-PAM} \end{bmatrix} = \begin{bmatrix} 2^{c-3} \\ 8^{c/2-1} \\ 8^{c/2-1}(M-1)^c\left(\log_2(M)\right)^{-c/2} \end{bmatrix} \tag{8}$$

and, [35, 46, 47]:

$$\begin{bmatrix} B_{RZ-OOK}(\mu) \\ B_{NRZ-OOK}(\mu) \\ B_{M-PAM}(\mu) \end{bmatrix} = \begin{bmatrix} c^2/\mu \\ 2c^2/\mu \\ 2\left[c(M-1)\right]^2/\left[\mu\log_2(M)\right] \end{bmatrix} \tag{9}$$

B. Time jitter influence on each individual link

The next studied issue of this work is about the error performance of an FSO link due to the time jitter effect. Supposing that the receiver detects the incoming signal pulse correctly at its center, it can be assumed that the misdetection probability is symmetrically distributed around the center of the current time slot, [14, 19]. Thus, by assuming T as the time instant of the pulse detection, the time jitter effect can be determined by a normal distribution with zero mean value, i.e. $\mu_T=0$ by supposing T=0 at the time slot center and variance σ_T^2, with the following PDF, [19, 49]:

$$f_T(T) = \frac{1}{\sqrt{2\pi}\sigma_T} \exp\left[-\frac{(T-\mu_T)^2}{2\sigma_T^2}\right] \tag{10}$$

The time slot duration, t_{sl}, is strongly related to the total bit rate, BR, and the used modulation scheme of the communication system. Thereinafter, the average BER of an FSO link, only due to the time jitter effect, is given as, [19]:

$$BER_{TJ} = c_s\left[\int_{-\infty}^{-t_{sl,s}} f_T(T)dT + \int_{t_{sl,s}}^{+\infty} f_T(T)dT\right] \tag{11}$$

for $c_{OOK} = 1/2$, $t_{sl,OOK} = 1/BR$ and $c_{PAM} = (1-1/M)\log_2(M)$, $t_{sl,PAM} = \log_2(M)/BR$, respectively, [18].

By solving the integrals of (11), the average BER for the OOK modulation scheme is given as:

$$BER_{TJ,OOK} = \frac{1}{4}\sum_{i=1}^{2} erfc\left(\frac{BR^{-1}+(-1)^i\mu_T}{\sqrt{2}\sigma_T}\right) \quad (12)$$

while for the M-PAM scheme the corresponding expression is given as:

$$BER_{TJ,PAM} = \frac{(M-1)m}{2M}\sum_{i=1}^{2} erfc\left(\frac{m/BR+(-1)^i\mu_T}{\sqrt{2}\sigma_T}\right) \quad (13)$$

where $m = \log_2 M$.

C. Joint ABER performance of each individual FSO link

Taking into account the above results, the average BER of an FSO link, under the joint action of atmospheric turbulence and time jitter effects, can be estimated. More specifically, the total bit error rate probability for each point to point (PtP) FSO link will be equal to the probability of a wrong decision at the receiver due to either a turbulence induced signal degradation or/and a jittered time shift misdetection. Thus, the total average BER of each FSO sub-link will be given as:

$$BER_s = BER_{Turb,s} + BER_{IJ,s} - BER_{Turb,s}BER_{TJ,s} \quad (14)$$

Applying (14) by appropriately combining (7)-(9) with (12) and (13) the joint average BER performance of an FSO link for any studied modulation scheme is obtained. Starting from the RZ-OOK case, the average BER is given as:

$$BER_{RZ-OOK} = 2^{c-3}C(\mu)\phi(\mu) + 2^{-2}\rho_{OOK} - 2^{c-5}C(\mu)\phi(\mu)\rho_{OOK} \quad (15)$$

where $\varphi(\mu) = G_{2,3}^{2,2}\left(\frac{c^2}{\mu}\middle|\begin{matrix}1-\frac{c}{2},\frac{1-c}{2}\\ 0,\frac{1}{2},-\frac{c}{2}\end{matrix}\right)$ and $\rho_{OOK} = \sum_{i=1}^{2} erfc\left(\frac{BR^{-1}+(-1)^i\mu_T}{\sqrt{2}\sigma_T}\right)$.

Then, the average BER for the NRZ-OOK scheme is:

$$BER_{NRZ-OOK} = \left(2\sqrt{2}\right)^{c-2}C(\mu)\chi(\mu) + 2^{-2}\rho_{OOK} - 2^{-2}\left(2\sqrt{2}\right)^{c-2}C(\mu)\chi(\mu)\rho_{OOK} \quad (16)$$

with $\chi(\mu) = \varphi(\mu/2)$.

As concerns the M-PAM scheme, the average BER is estimated through the next formula:

$$BER_{PAM} = \left(2\sqrt{2}\right)^{c-2}(M-1)^c m^{-\frac{c}{2}}C(\mu)\psi(\mu) + (M-1)(2M)^{-1}m\rho_{PAM} - \left(2\sqrt{2}\right)^{c-2}\times (M-1)^{c+1}(2M)^{-1}m^{1-\frac{c}{2}}C(\mu)\psi(\mu)\rho_{PAM} \quad (17)$$

where: $\psi(\mu) = G_{2,3}^{2,2}\left(\frac{2[(M-1)c]^2}{\mu\log_2 M}\middle|\begin{matrix}1-\frac{c}{2},\frac{1-c}{2}\\ 0,\frac{1}{2},-\frac{c}{2}\end{matrix}\right)$ and $\rho_{PAM} = \sum_{i=1}^{2} erfc\left(\frac{m/BR+(-1)^i\mu_T}{\sqrt{2}\sigma_T}\right)$.

D. Total ABER of the whole DF relayed FSO link

Finally, the presence of a number of serially DF relayed nodes between the initial transmitter and the final receiver is studied in order to derive the total average BER of the optical communication system. Assuming an aggregate link consisting of L individual FSO links, connected through DF relays, the following formula gives the total average BER of the end-to-end link, [50, 51]:

$$BER_{total} = \sum_{i=1}^{L}\left[BER_i\prod_{j=i+1}^{L}\left(1-2BER_j\right)\right] \quad (18)$$

In particular, supposing same modulation scheme for each individual link operation, the average BER of the total link using RZ-OOK is given in (19) by simply applying (15) into (18), while for the NRZ-OOK case the expression (20) is obtained by substituting (16) into (18) and for the PAM scheme, Eq. (21) arises by applying (17) into (18).

$$BER_{RZ-OOK,total} = \sum_{i=1}^{L}\left[\left(2^{c_i-3}C_i(\mu)\phi_i(\mu) + 2^{-2}\rho_{OOK,i} - 2^{c_i-5}C_i(\mu)\phi_i(\mu)\rho_{OOK,i}\right)\times \prod_{j=i+1}^{L}\left(1-2^{c_j-2}C_j(\mu)\phi_j(\mu) - 2^{-1}\rho_{OOK,j} + 2^{c_j-4}C_j(\mu)\phi_j(\mu)\rho_{OOK,j}\right)\right] \quad (19)$$

$$BER_{NRZ-OOK,total} = \sum_{i=1}^{L}\left\{\left[\left(2\sqrt{2}\right)^{c_i-2}C_i(\mu)\chi_i(\mu) + 2^{-2}\rho_{OOK,i} - 2^{-2}\left(2\sqrt{2}\right)^{c_i-2}C_i(\mu)\chi_i(\mu)\rho_{OOK,i}\right]\times \prod_{j=i+1}^{L}\left[1-2\left(2\sqrt{2}\right)^{c_j-2}C_j(\mu)\chi_j(\mu) - 2^{-1}\rho_{OOK,j} + 2^{-1}\left(2\sqrt{2}\right)^{c_j-2}C_j(\mu)\chi_j(\mu)\rho_{OOK,j}\right]\right\} \quad (20)$$

$$BER_{M-PAM,total} = \sum_{i=1}^{L}\left\{\left[\left(2\sqrt{2}\right)^{c_i-2}(M-1)^{c_i}m^{-\frac{c_i}{2}}C_i(\mu)\psi_i(\mu) + (M-1)(2M)^{-1}m\rho_{PAM,i} - \left(2\sqrt{2}\right)^{c_i-2}(M-1)^{c_i+1}(2M)^{-1}m^{1-\frac{c_i}{2}}C_i(\mu)\psi_i(\mu)\rho_{PAM,i}\right]\times \prod_{j=i+1}^{L}\left[1-2\left(2\sqrt{2}\right)^{c_j-2}(M-1)^{c_j}m^{-\frac{c_j}{2}}C_j(\mu)\psi_j(\mu) - (M-1)M^{-1}m\rho_{PAM,j} + \left(2\sqrt{2}\right)^{c_j-2}(M-1)^{c_j+1}M^{-1}m^{1-\frac{c_j}{2}}C_j(\mu)\psi_j(\mu)\rho_{PAM,j}\right]\right\} \quad (21)$$

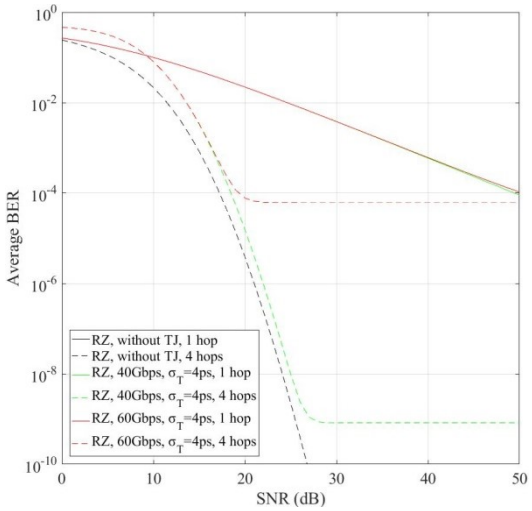


Fig. 1. ABER as a function of SNR for single and quad-hop links with or without the influence of time jitter effect for two bit rate values and σ_T=4ps (λ=1.55 μm, $C_n^2 = 5\times10^{-15}\,m^{-2/3}$), using RZ-OOK modulation scheme.

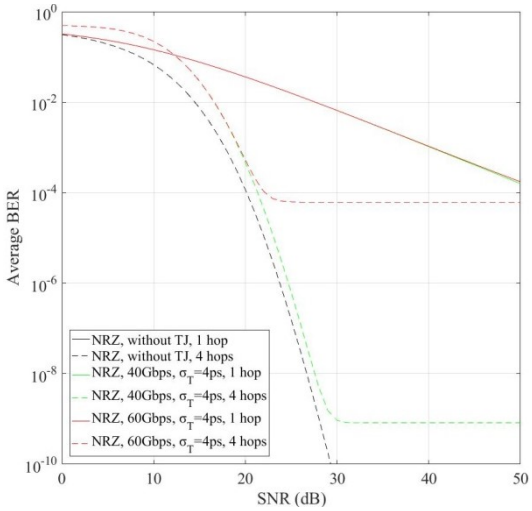


Fig. 2. ABER as a function of SNR for single and quad-hop links with or without the influence of time jitter effect for two bit rate values and σ_T=4ps (λ=1.55 μm, $C_n^2 = 5\times10^{-15}\,m^{-2/3}$), using NRZ-OOK modulation scheme.

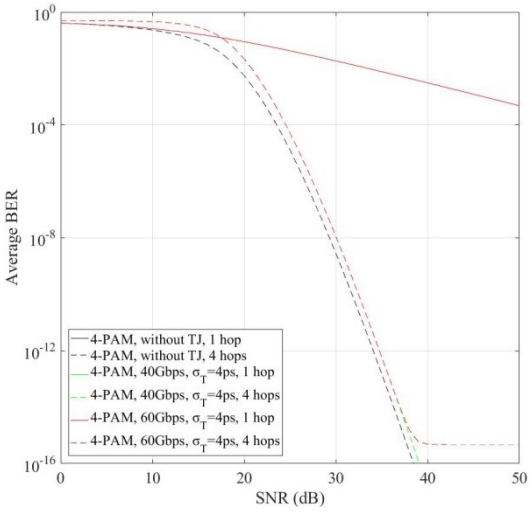


Fig. 3. ABER as a function of SNR for single and quad-hop links with or without the influence of time jitter effect for two bit rate values and σ_T=4ps (λ=1.55 μm, $C_n^2 = 5\times10^{-15}\,m^{-2/3}$), using 4-PAM scheme.

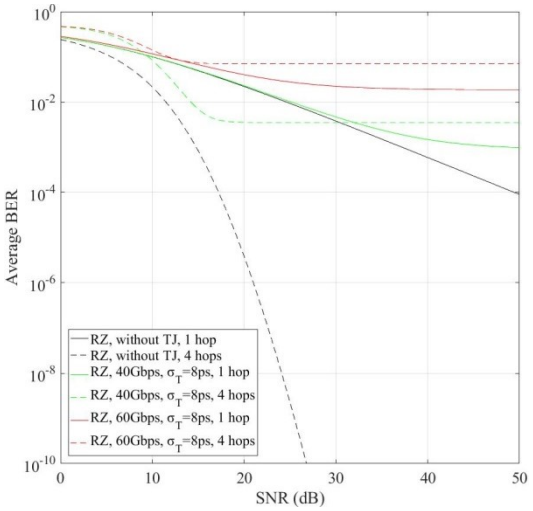


Fig. 4. ABER as a function of SNR for single and quad-hop links with or without the influence of time jitter effect for two bit rate values and σ_T=8ps (λ=1.55 μm, $C_n^2 = 5\times10^{-15}\,m^{-2/3}$), using RZ-OOK modulation scheme.

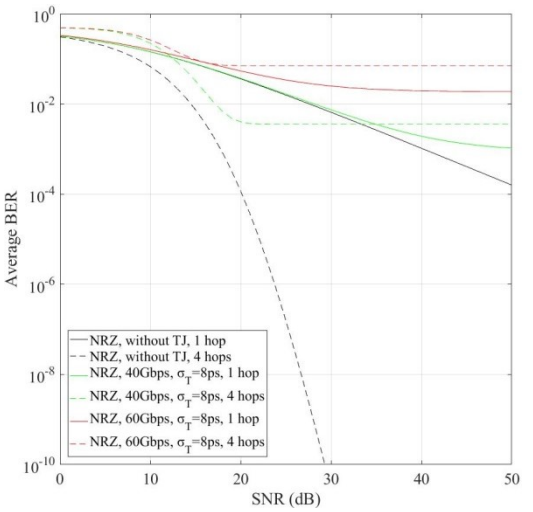


Fig. 5. ABER as a function of SNR for single and quad-hop links with or without the influence of time jitter effect for two bit rate values and σ_T=8ps (λ=1.55 μm, $C_n^2 = 5\times10^{-15}\,m^{-2/3}$), using NRZ-OOK modulation scheme.

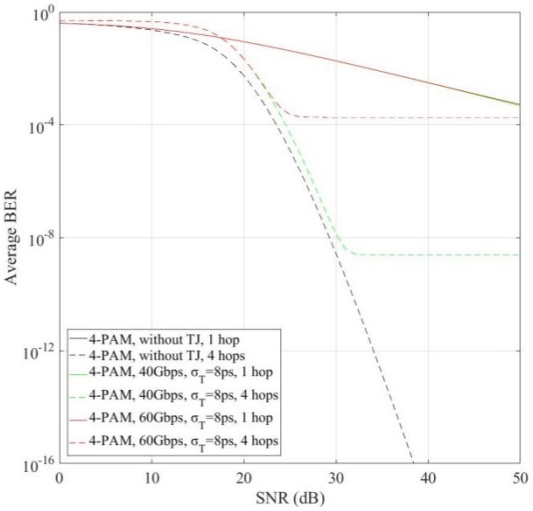


Fig. 6. ABER as a function of SNR for single and quad-hop links with or without the influence of time jitter effect for two bit rate values and σ_T=8ps (λ=1.55 μm, $C_n^2 = 5\times10^{-15}\,m^{-2/3}$), using 4-PAM scheme.

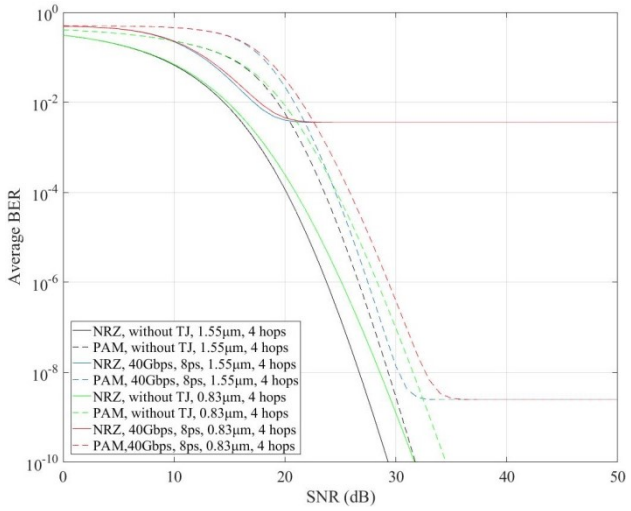


Fig. 7. ABER as a function of SNR for quad-hop links with or without the influence of time jitter effect for two wavelength values, 40 Gbps and σ_T=8ps ($C_n^2 = 5\times10^{-15}$ m$^{-2/3}$), using NRZ-OOK and 4-PAM schemes.

III. Numerical Results

In the previous section we derived analytically the appropriate expressions for the performance estimation of the FSO links under the action of weak turbulence and time jitter effects. Here, the average BER results are presented for FSO parameter values which have been chosen in order to plainly present the influence of each effect at the link's performance. It should be mentioned here using the expressions (19)-(21), the average BER performance can be estimated for any parameter values corresponding to each specific FSO communication system.

In particular, the diameter of the receiver's aperture is set to 0.1 m, the operating wavelength of the system and the value of the structure index parameter, C_n^2, in Figs 1-6 are fixed to 1.55 μm and 5×10^{-15} m$^{-2/3}$, respectively. Also, the wavelength of 0.83 μm and the condition for structure index parameter value of 1×10^{-15} m$^{-2/3}$ are alternatively tested in Fig. 7 and Fig. 8, respectively, in order to clearlypresent the wavelength and the turbulence influence at the link's performance. Regarding the time jitter parameters, the mean value of the normal distribution is set to zero for all the cases, whereas the standard deviation parameter takes two values, i.e. 4 ps and 8 ps.

In all cases, the simulations are performed for bit rate values either of 40 Gbps or 60 Gbps, while the total link length has been assumed to be 8 km, implemented with single or quadruple hops, meaning, for the latter case, that each individual link length becomes equal to the quarter of the total link distance. The specific data rates have been chosen in order to present plainly the influence of time jitter effect at the link's performance. The conditions described above are applied to OOK and PAM schemes variants, and their performance is depicted separately in Figs 1-6.

In particular, Fig. 1 depicts the average BER performance curves as a function of the SNR in dB for any combination of either one or four hops, without or with time jitter effect with σ_T=4 ps and bitrate of 40 Gbps or 60 Gbps, for RZ-OOK modulation scheme. The corresponding curves for NRZ-OOK and 4-PAM are shown in Figs 2 and 3, respectively. To demonstrate the influence of time jitter's strength, the corresponding outcomes are obtained for σ_T=8 ps in Figs 4, 5 and 6, respectively. As mentioned above, alternative values of wavelength and C_n^2 for the quadruple hops case, NRZ-OOK and 4-PAM schemes are illustrated in Fig. 7 and Fig. 8, respectively.

At first glance, two predominant elements are present in the figures. At first, the time jitter effect sets saturation thresholds in ABER performance, beyond which no further average BER improvement can be achieved and whose values depend on the bit rate and the strength of the time jitter effect, σ_T. Secondly, the DF relayed multi-hop effect, which allows further improvement of BER dependence, due to the steeper slope of respective curve, at a cost of higher thresholds, for any modulation scheme. Concerning the time jitter effect, the presented results are quite reasonable, because ABER performance is highly affected by the bit rate and the standard deviation values, since both higher rates, meaning shorter detection time slots, and higher temporal deviations, which become more comparable to the bit's time slot, generate stronger time jitter influence and poorer ABER performances, consequently.

In addition, although the use of multiple DF relays either optimizes the reliability, i.e. lower average BER performance for equal SNR, or minimizes the power requirements along links, i.e. equal average BER outcomes for smaller SNR values, multi-hop deteriorates the accumulated BER performance, which is limited by higher and higher thresholds as the number of hops increases. So, the net benefit of using a number of DF relays is a tradeoff between the desirable improvement and the loss induced. Moreover, any expected improvement may be eventually canceled due to strong time jitter effects. Thereafter, the modulation scheme used plays a key role in terms of BER performance.

In total, the OOK variants seem to be more vulnerable than 4-PAM counterpart. With stronger time jitter effect, it is straightforward that OOK variants are rather prohibitive for high-rate communications, because of the extremely poor BER values performed at such rates. On the other hand, compared to OOK, 4-PAM is a more suitable scheme for high rate communications, with better BER performance. Also, it is worth noticing that the ABER thresholds regarding OOK are identical and dependent only on hop number and bit rate, regardless of the variant scheme.

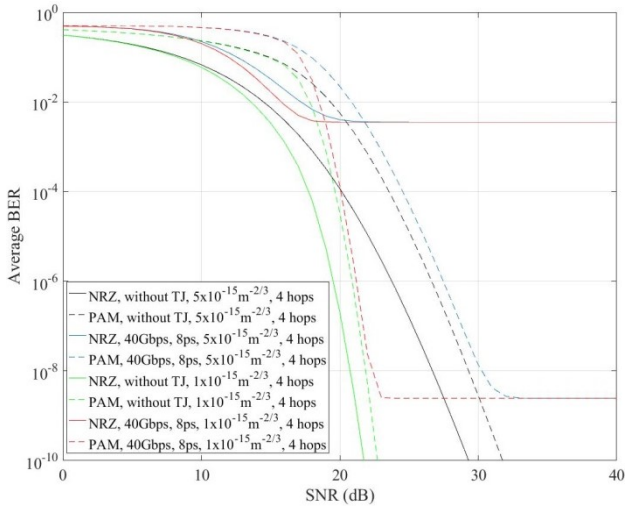


Fig. 8. ABER as a function of SNR for quad-hop links with or without the influence of time jitter effect for two C_n^2 values, 40 Gbps and σ_T=8ps (λ=1.55 μm), using NRZ-OOK and 4-PAM schemes.

As regards the variation of the wavelength in Fig. 7, it seems that the cases with the larger value, i.e. 1.55 μm, achieves better performance results, before the saturation point which depends mainly on the time jitter effect. On the other hand, the curves of Fig. 8 show that the turbulence effect affects strongly the link's performance. Thus, for weaker turbulence, the average BER of the link is much better than for stronger, in the area before the saturation point.

IV. Conclusions

The scope of this work was the study, for the first time to the best of our knowledge, of the joint effects of atmospheric turbulence and time jitter on serially DF relayed FSO links. New closed form expressions for the estimation of the average BER have been derived, while the corresponding numerical results using typical parameter values have been obtained and several important and applicable conclusions have been extracted. Conclusively, it can be deduced that the apparent superiority of the multi-hop approach concerning the average BER performance is more or less overshadowed by the strong or very strong presence of time jitter effect. Furthermore, regardless of the operating wavelength and the prevailing turbulence conditions, saturation thresholds beyond which no further average BER improvement can be achieved are identically appeared, varying with the bit rate, the time jitter variance, the number of hops, and the modulation scheme selected.

Acknowledgment

H.E. Nistazakis, P.J. Gripeos and G.S. Tombras acknowledge funding from National and Kapodistrian University of Athens, Special Account for Research Grants. G.D. Roumelas acknowledges that this research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Program "Human Resources Development, Education and Lifelong Learning" in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

References

[1] M. A. Khalighi and M. Uysal, "Survey on free space optical communication: A communication theory perspective," IEEE Commun. Surv. Tutorials, vol. 16, no. 4, pp. 2231–2258, 2014.

[2] F. Demers, H. Yanikomeroglu, and M. St-Hilaire, "A survey of opportunities for free space optics in next generation cellular networks," Proc. - 2011 9th Annu. Commun. Networks Serv. Res. Conf. CNSR 2011, pp. 210–216, 2011.

[3] M. Uysal, C. Capsoni, Z. Ghassemlooy, A. Boucouvalas, E. Udvary, Optical Wireless Communications. Cham: Springer International Publishing, 2016.

[4] Z. Ghassemlooy and W. O. Popoola, "Terrestrial Free-Space Optical Communications," in Mobile and Wireless Communications Network Layer and Circuit Level Design, InTech, 2010.

[5] L. C. Andrews, R. L. Phillips, and C. Y. Young, Laser Beam Scintillation with Applications. SPIE, 2001.

[6] M. S. Awan, L. C. Horwath, S. S. Muhammad, E. Leitgeb, F. Nadeem, and M. S. Khan, "Characterization of Fog and Snow Attenuations for Free-Space Optical Propagation," J. Commun., vol. 4, no. 8, pp. 533–545, Sep. 2009.

[7] M. H. Ibrahim, H. A. Shaban, and M. H. Aly, "Effect of different weather conditions on BER performance of single-channel free space optical links," Optik (Stuttg)., vol. 137, pp. 291–297, May 2017.

[8] A. Chaman-Motlagh, V. Ahmadi, and Z. Ghassemlooy, "A modified model of the atmospheric effects on the performance of FSO links employing single and multiple receivers," J. Mod. Opt., vol. 57, no. 1, pp. 37–42, Jan. 2010.

[9] W. O. Popoola, Z. Ghassemlooy, H. Haas, E. Leitgeb, and V. Ahmadi, "Error performance of terrestrial free space optical links with subcarrier time diversity," IET Commun., vol. 6, no. 5, p. 499, 2012.

[10] M. A. Al-Habash, "Mathematical model for the irradiance probability density function of a laser beam propagating through turbulent media," Opt. Eng., vol. 40, no. 8, p. 1554, Aug. 2001.

[11] M. S. Ahmed and T. Gucluoglu, "Performance of generalized frequency division multiplexing over gamma gamma free space optical link," Opt. Commun., vol. 466, no. February, p. 125683, Jul. 2020.

[12] N. Lazer and Y. P. Arul Teen, "Free Space Optical Communication and Laser Beam Propagation through Turbulent Atmosphere: A Brief Survey," 2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC), Nagercoil, India, 2019, pp. 1-6, doi: 10.1109/ICRAECC43874.2019.8994973.

[13] S. Dimitrov, S. Sinanovic, and H. Haas, "Clipping Noise in OFDM-Based Optical Wireless Communication Systems," IEEE Trans. Commun., vol. 60, no. 4, pp. 1072–1081, Apr. 2012.

[14] G. P. Agrawal, Fiber-Optic Communication Systems, vol. 6. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010.

[15] M. J. Underhill, "Time jitter and phase noise - Now and in the future?," in 2012 IEEE International Frequency Control Symposium Proceedings, 2012, pp. 1–8.

[16] Z. Ghassemlooy, S. Arnon, M. Uysal, Z. Xu, and J. Cheng, "Emerging Optical Wireless Communications-Advances and Challenges," IEEE J. Sel. Areas Commun., vol. 33, no. 9, pp. 1738–1749, 2015.

[17] V. S. Grigoryan, C. R. Menyuk, and R. M. Mu, "Calculation of timing and amplitude jitter in dispersion-managed optical fiber communications using linearization," J. Light. Technol., vol. 17, no. 8, pp. 1347–1356, 1999.

[18] A. H. Gnauck, A. Mecozzi, C. B. Clausen, Sang-Gyu Park, and M. Shtaif, "Cancellation of timing and amplitude jitter in symmetric links using highly dispersed pulses," IEEE Photonics Technol. Lett., vol. 13, no. 5, pp. 445–447, 2002.

[19] G. D. Roumelas, H. E. Nistazakis, A. N. Stassinakis, G. K. Varotsos, A. D. Tsigopoulos, and G. S. Tombras, "Time Jitter, Turbulence and Chromatic Dispersion in Underwater Optical Wireless Links," Technologies, vol. 8, no. 1, p. 3, Dec. 2019.

[20] H. E. Nistazakis, D. J. Frantzeskakis, J. Atai, B. A. Malomed, N. Efremidis, and K. Hizanidis, "Multichannel pulse dynamics in a stabilized Ginzburg-Landau system," Physical Review E, Vol. 65, 036605, 2002.

[21] Y. Wang, Y. Zhang, Z. Dou and D. Tian, "The Influence of Timing Error on the Performance of Optical Pulse PPM System in Atmospheric Turbulent Channels," 2010 Symposium on Photonics and Optoelectronics, Chengdu, 2010, pp. 1-4, doi: 10.1109/SOPO.2010.5504021.

[22] K. M. N. Islam and S. P. Majumder, "Effect of timing jitter on the BER performance of a M-PPM FSO link over atmospheric turbulence channel," 8th International Conference on Electrical and Computer Engineering, Dhaka, 2014, pp. 409-412, doi: 10.1109/ICECE.2014.7027017.

[23] Y. Li, T. Geng, S. Ma, S. Gao, H. Gao, "Timing jitter's influence on the symbol error rate performance of the L-ary pulse position modulation free-space optical link in atmospheric turbulent channels with pointing errors," Optical Engineering 56(3) 036116 (2017). https://doi.org/10.1117/1.OE.56.3.036116.

[24] D. G. Brennan, "Linear Diversity Combining Techniques," in Proceedings of the IRE, vol. 47, no. 6, pp. 1075-1102, June 1959, doi: 10.1109/JRPROC.1959.287136.

[25] R. W. Hamming, "Error detecting and error correcting codes," in The Bell System Technical Journal, vol. 29, no. 2, pp. 147-160, April 1950, doi: 10.1002/j.1538-7305.1950.tb00463.x.

[26] B. Epple, "Simplified Channel Model for Simulation of Free-Space Optical Communications," J. Opt. Commun. Netw., vol. 2, no. 5, p. 293, 2010.

[27] H. G. Sandalidis, "Performance Analysis of a Laser Ground-Station-to-Satellite Link With Modulated Gamma-Distributed Irradiance Fluctuations," J. Opt. Commun. Netw., vol. 2, no. 11, p. 938, 2010.

[28] G. K. Varotsos, H. E. Nistazakis, C. K. Volos, and G. S. Tombras, "FSO links with diversity pointing errors and temporal

broadening of the pulses over weak to strong atmospheric turbulence channels," Optik (Stuttg)., vol. 127, no. 6, pp. 3402–3409, 2016.

[29] W. Gappmair, "Novel results on pulse-position modulation performance for terrestrial free-space optical links impaired by turbulent atmosphere and pointing errors," IET Commun., vol. 6, no. 10, p. 1300, 2012.

[30] H. G. Sandalidis and T. A. Tsiftsis, "Outage probability and ergodic capacity of free-space optical links over strong turbulence," Electron. Lett., vol. 44, no. 1, p. 46, 2008.

[31] S. Dimitrov, S. Sinanovic, and H. Haas, "Signal Shaping and Modulation for Optical Wireless Communication," J. Light. Technol., vol. 30, no. 9, pp. 1319–1328, May 2012.

[32] G. Baiden, Y. Bissiri, and A. Masoti, "Paving the way for a future underwater omni-directional wireless optical communication systems," Ocean Eng., vol. 36, no. 9–10, pp. 633–640, 2009.

[33] A. N. Stassinakis, H. E. Nistazakis, and G. S. Tombras, "Comparative performance study of one or multiple receivers schemes for FSO links over gamma–gamma turbulence channels," J. Mod. Opt., vol. 59, no. 11, pp. 1023–1031, Jun. 2012.

[34] M. Uysal, J. Li, and M. Yu, "Error rate performance analysis of coded free-space optical links over gamma-gamma atmospheric turbulence channels," IEEE Trans. Wirel. Commun., vol. 5, no. 6, pp. 1229–1233, 2006.

[35] A. N. Stassinakis, H. E. Nistazakis, K. P. Peppas, and G. S. Tombras, "Improving the availability of terrestrial FSO links over log normal atmospheric turbulence channels using dispersive chirped Gaussian pulses," Opt. Laser Technol., vol. 54, no. 4, pp. 329–334, Dec. 2013.

[36] G. T. Djordjevic, M. I. Petkovic, A. M. Cvetkovic, and G. K. Karagiannidis, "Mixed RF/FSO Relaying With Outdated Channel State Information," IEEE J. Sel. Areas Commun., vol. 33, no. 9, pp. 1935–1948, Sep. 2015.

[37] A. K. Majumdar, Advanced Free Space Optics (FSO) - A System Approach, vol. 140. 2015.

[38] W. Gappmair and H. E. Nistazakis, "Subcarrier PSK Performance in Terrestrial FSO Links Impaired by Gamma-Gamma Fading, Pointing Errors, and Phase Noise," J. Light. Technol., vol. 35, no. 9, pp. 1624–1632, 2017.

[39] Xiaoming Zhu and J. M. Kahn, "Free-space optical communication through atmospheric turbulence channels," IEEE Trans. Commun., vol. 50, no. 8, pp. 1293–1300, Aug. 2002.

[40] T. A. Tsiftsis, H. G. Sandalidis, G. K. Karagiannidis, and M. Uysal, "Optical wireless links with spatial diversity over strong atmospheric turbulence channels," IEEE Trans. Wirel. Commun., vol. 8, no. 2, pp. 951–957, Feb. 2009.

[41] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. 1965.

[42] H. Weber and G. B. Arfken, Essentials of Math Methods for Physicists. Elsevier, 1966.

[43] K. Prabu, D. S. Kumar, and T. Srinivas, "Performance analysis of FSO links under Strong atmospheric turbulence conditions using various modulation schemes," Optik (Stuttg)., vol. 125, no. 19, pp. 5573–5581, 2014.

[44] X. Yi, Z. Liu, P. Yue, and T. Shang, "BER Performance Analysis for M-ary PPM over Gamma-Gamma Atmospheric Turbulence Channels," 2010 Int. Conf. Comput. Intell. Softw. Eng., pp. 1–4, Sep. 2010.

[45] A. Dang, "A closed-form solution of the bit-error rate for optical wireless communication systems over atmospheric turbulence channels," Opt. Express, vol. 19, no. 4, p. 3494, 2011.

[46] Z. Ghassemlooy, W. O. Popoola, S. Gao, J. I. H. Allen, and E. Leitgeb, "Free-space optical communication employing subcarrier modulation and spatial diversity in atmospheric turbulence channel," IET Optoelectron., vol. 2, no. 1, pp. 16–23, Feb. 2008.

[47] J. G. Proakis and M. Salehi, Communication Systems Engineering, 2nd ed. New Jersey: Prentice-Hall, 2002.

[48] V. S. Adamchik and O. I. Marichev, "The algorithm for calculating integrals of hypergeometric type functions and its realization in REDUCE system," in Proceedings of the international symposium on Symbolic and algebraic computation - ISSAC '90, 1990, pp. 212–224.

[49] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd ed. United States of America: McGraw-Hill, Inc., 1991.

[50] M. Sheng, P. Jiang, Q. Hu, Q. Su, and X. X. Xie, "End-to-end average BER analysis for multihop free-space optical communications with pointing errors," J. Opt. (United Kingdom), vol. 15, no. 5, 2013.

[51] E. Morgado, I. Mora-Jimenez, J. J. Vinagre, J. Ramos, and A. J. Caamano, "End-to-End Average BER in Multihop Wireless Networks over Fading Channels," IEEE Trans. Wirel. Commun., vol. 9, no. 8, pp. 2478–2487, Aug. 2010.

Implementation of Experimental Modulator into the Luminaire of Public Lighting Based on the OOK Modulation with Bias-Tee

Stanislav Hejduk, Tomas Stratil, Jan Latal, Ales Vanderka, Lukas Hajek, Jakub Kolar
VSB–Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science,
Department of Telecommunications, 17. listopadu 15, Ostrava, 708 33, Czech Republic
stanislav.hejduk@vsb.cz, tomas.stratil@vsb.cz, jan.latal@vsb.cz, ales.vanderka@vsb.cz,
lukas.hajek@vsb.cz, jakub.kolar@vsb.cz

***Abstract*—This article brings new possibilities of using luminaires of public lighting with the option of implementing an experimental modulator based on the OOK with Bias-Tee to increase functionality and possibility of using the public lighting network within intravilanes. In the article, the block diagrams of the OOK modulator for the transmitting part with Bias-Tee including the block diagram for the receiving part are introduced. Other parts of the article focus attention on verification of the functionality of the proposed concept with the aim of achieving signal transmission for road users through the luminaires of public lighting.**

***Index Terms*—Transmitter, Receiver, baud rate, OOK, Bias-Tee, Public Lighting, VLC.**

I. Introduction & Related Works

Nowadays we can witness a sharp increase in data transferred to end stations or customers as the possibility of networking sites from mobile operators or Internet provides have been increasing. This, however, presents new challenges in the form of the transmission of a loss-less signal, if possible, and thus the question of where to place the transmitting or, more precisely, receiving unit. One way that would solve this problem is to use a network of public lighting poles. It is known that any construction is related to complex administrative and legislative acts, which considerably increase the cost and extend the time of possible construction [1], [2].

As the IoT (Internet of Things) concepts or SMART elements evolve, the possibility of using a number of tools and sensor data for transmission to road users go hand in hand. The concept of the number of tools presents new possibilities of solutions for intersection control systems, traffic optimization based on utilisation by road users, transmission of important information /telematics data.

Someone accesses the SMART lighting solution through the use of Raspberry-Pi 3 with a Wi-Fi module combination [3].

Another possibility is to use a matrix of LED (Light-Emitting Diode) radiation sources and their modulation by means of the OOK (On-Off Keying) through the FPGA (Field Programmable Gate Array) field, where 40 meters of communication distance was achieved for low transmission rates with background noise [4].

The possibilities of using the FPGA fields for construction of the transmitter prototype on the basis Li-Fi for the real-time communication were examined. It was incorporated into the luminaire of the PL (Public Lighting) by the transmitter with the aim of achieving qualitative parameters of communication and lighting simultaneously [5].

Other authors, with the aim of transmitting ITS (Intelligent Transportation Systems) data, focused on the direct implementation of the transmitter and receiver of the front lights of cars through VLC (Visible Light Communication) to reach Q-factor 5 and BER=10^9. The standard CAN (Controller Area Network) bus applied for cars was used for the control of their prototypes [6].

Considering the new possibilities that bring the use of semiconductor radiation sources, there is also the possibility of dimming with suppression of the function of flashing luminaires by means of a suitable type of modulation format and control of the PL network. For example, the DFSOOK (Dimming frequency shift On-Off keying) based on the FSK (Frequency Shift Keying) is best suited for incoherent detection of signal component.

A scheme demonstrating the use of the DFSOOK for the VLC with PWM (Pulse Width Modulation) dimming control was proposed. The transmitter has been tested for low error rate (BER) and longer rage of communication and lighting [7].

Our objective of the work is to design the transmitter and receiver based on the OOK modulation with our design of Bias-Tee circuit for luminaires of public lighting with the achievement of useful signal information. The article presents possibilities of the proposal of block diagrams and measurements of signal recording at setting different modulation rates for data transmission.

II. OOK modulator with Bias-Tee

In general, it is not possible to transmit the data directly for the visible spectrum and OOK by public lighting. Simple switching according to logical levels would result in a noticeable variation on the light intensity. Switching high power public lighting has big drawbacks like switching speed and light intensity control. Based on that, the Bias-Tee circuit

[8], [9] was used. In addition, when using a Bias-Tee circuit, a minimum operating frequency must be kept so that longer sequences of values of the same level could present a problem on the transmitting side.

The first task is to modify the data into a format that can be sent by means of the Bias-Tee circuit (see Fig. 1). Thus, it is important to select a suitable coding scheme, such as a differential Manchester, or any other solution that will maintain a stable mean value of the signal and a minimum operating frequency for transmission. Nevertheless, within the experiments, the implementation of similar schemes. The data is modulated to a carrier signal of higher frequency and adjusted to a symmetrical form. The advantage of this solutions is that it is possible to use the original power supply of the luminaire of public lighting and, in case of failure of the communication part, the lighting function of the public lighting will not be interrupted. This is very important for the operators and organisational entities of public networks within intravilanes.

III. OOK MODULATION (ON-OFF KEYING) OF THE POWER LED

The OOK modulation belongs to the simplest types where the logical value 1 is coded as a light pulse. In order to reduce the complexity of the modulator the pulses with a rectangular shape are used. The baud rate (R_b) of one bit is presented as:

$$R_b = \frac{1}{T_b}, \tag{1}$$

where T_b indicated duration of one bit. An important parameter (except BER), which is necessary to consider in each modulation scheme is the bandwidth requirement. The bandwidth is estimated by the spectral density of the signal obtained by the Fourier transformation, using the auto-correlation function. The spectral density of the OOK modulated signal without input correlation has the form of [10]:

$$S(f)_{OOK} = \frac{i_s^2}{4R_b} sinc^2(\frac{\pi f}{R_b}) \left[1 + R_b \sum_{k=-\infty}^{\infty} \delta(f - kR_b)\right], \tag{2}$$

where: $\mathrm{sinc}(x) = \sin(x)/x$, δ(x) Dirac function, i_s average value of photoelectric current generated in the optical radiation source, f frequency. Since the duration of the pulse is finite, the spectrum extends to infinity. A zero-frequency pulse corresponds to the DC component and represents the energy balance. As the pulse value δ decreases, the need of bandwidth increases. If the value is δ = 0.5 we usually call this modulation scheme as OOK RZ (with a return to zero). It can increase the bandwidth up to twofold compared to above mentioned modulation scheme OOK NRZ (with a non-return to zero).

IV. RECEIVING PART OF THE SYSTEM

The receiver (see Fig. 2) has the task to receive the signal even in inhospitable conditions of daylight when the signal from the luminaire public lighting is disproportionate to the probable interference of direct solar radiation to a large extent. The supply voltage must be compatible with the 12 V on-board voltage of a vehicle. Therefore, only +/-6 V is available on the receiving side. In addition, there is a condition of a minimum modulation frequency of 1 MHz during detection, the photodetector must not be saturated with parasitic signal sources.

Therefore, the initial transimpedance amplifier must be set to be able to receive data even in direct sunlight. Part of the problem can be solved by filtering unnecessary portions of the signal spectrum (we keep only blue colour for communication. However, setting a higher sensitivity results has a negative impact in the form of limiting the maximum operating frequencies.

After filtering, the signal is amplified so that the signal values would be as high as possible (saturation of the signal is directly desirable in this case) but the noise must not be higher than the decision level for signal processing. As a result, the system is capable of operating in a very large dynamic range and is not limited by the need of evaluating levels according to the signal amplitude.

The demodulator block has the task to transform the signal into a data stream compatible with the evaluation unit. The received data are then processed based on current needs in this case using a web application. Nevertheless, it is not a problem to transmit data e.g. to the control unit or vehicle infotainment [**?**].

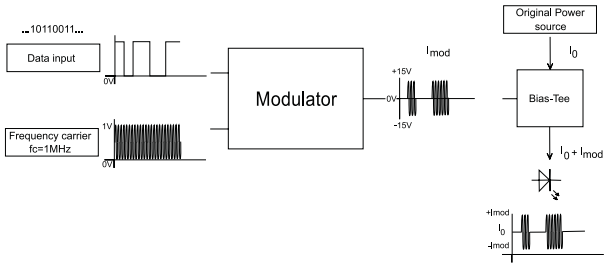


Fig. 1. Block diagram of the transmitting part.

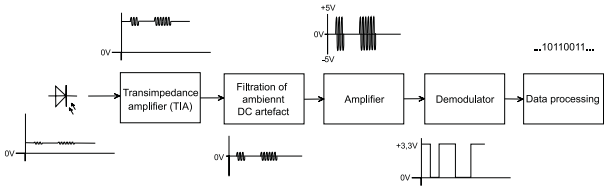


Fig. 2. Block diagram of the receiving part.

Figure 3 shows public lightning implementation of proposed solution of VLC experimental modulator. Connection maintain original power source for LED matrix and independent power source for modulator and control circuits. Wi-Fi and WEB interface allows complex control of light and communication functionality. DALI interface mediates light control like dimming and on/off functions. Experimental modulator mediates data adjusting for Bias-Tee and Bias-Tee part takes care of

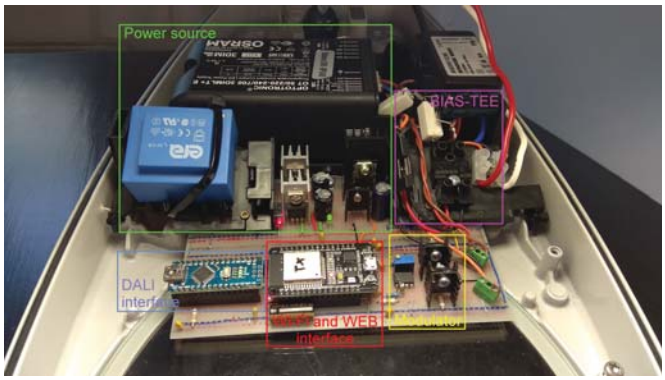


Fig. 3. Real implementation of proposed solution (transmitter part).

additional filtering of original power source and LED matrix modulation.

V. Setting Options

A web interface for the receiving and transmitting part was designed for the need of real-time experimentation (see Fig. 4). The transmitting part of the web interface allows the user to select the text of the message, baud rate and modulation frequency. And together with control of original DALI power source of the light (on/off/dimming) [11], [12]. The same baud rate can be simply set for the receiving part. Then, the received data that is transmitted through the public lighting will automatically display. The optimum system setup

Fig. 4. The web visualization to control and set baud rates and data for the modulator with Bias-Tee.

is adapted for the 1 MHz carrier frequency and the baud rate of 115.2 kBaud (see Fig. 5). Wireless control of transmitter was done by the HTTP server, HTTP Get, and Post methods. Due to testing in real-time WebSocket instance was created on the receiver site module. Both devices act as a Wi-Fi access point, no other network component is needed.

Theoretically, it is possible to reach up to 512 kBaud (see Fig. 6) with the proposed system at higher signal levels. However, it is already evident that the signal is no longer able to fully adapt to the transmitted one. Therefore, the electronics and appropriate setting of decision levels will come next. In

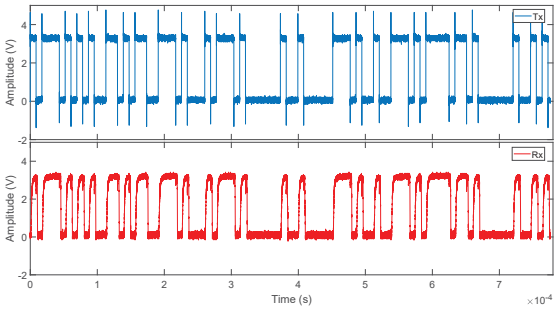


Fig. 5. Communication at 115 kBaud.

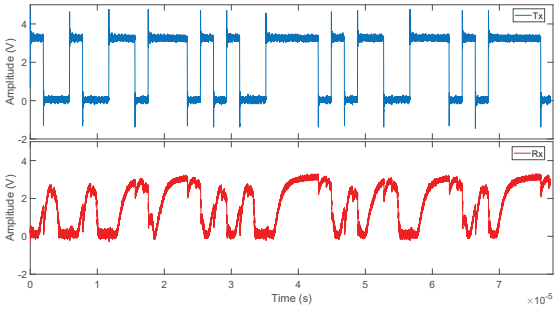


Fig. 6. Course of communication at 512 kBaud.

case of a weak signal (see Fig. 7), even if a high signal amplification is not sufficient, the amplitude of the signal is reduced. If, in this case, the decision level of the evaluation electronics is not exceeded, the data will be lost. A similar

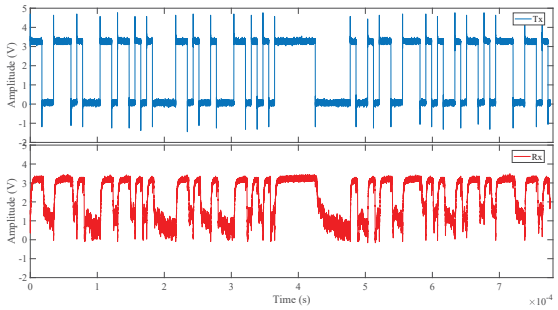


Fig. 7. Course of communication at a weak signal.

consequence will result in the selection of an unnecessarily high carrier frequency as the amplification gain decreases with increasing frequency. In combination with low power levels of the signal we can easily reach the boundaries in MHz units. A few cascaded amplifiers with bandwidth of 100 MHz were used for the construction when the amplification of individual stages could not be higher than 15.

Measured bandwidth limit for tested transmitter was over 3 MHz, which is a little bit higher than chosen 1 MHz. But there are few reasons for slow down (and all of them are connected to the receiver). First reason is photodiode–compromise between speed and sensitivity is ruthless, because

higher frequencies require higher incidental power for reliable detection. Second reason is the rest of the photodetector circuits–since the signal amplitude is quite low and noise is high, the signal amplification and recovery is easier at lower frequencies.

VI. Real deployment testing

First tests during the daylight (see Fig. 8) wasn't entirely successful. Direct sunlight easily overloads photodetector in any configuration (since VLC uses visible light spectrum). Even experiment with blue filter element didn't help. After few upgrades, the surrounding light tolerance of the photodetector was over 7000 lux, which was enough for cloudy weather. However, for sunny day still not enough. Second limitation was light dimming support. To keep the original dimming functionality, the depth of modulation was set only to 50 %. As a result, the communication range is almost constant within the reasonable dimming intensity. On the other hand, the covered area is few meters smaller then was expected.

VII. Conclusion

The communication system based on the public lighting is able to easily transmit data to road users. The limitations in terms of baud rates are largely dependent on the photodetector. The essence of the system is the rapid transmission of telemetric information, for instance, when a vehicle is passing under a public lighting luminaire. The potential of using such a system is not only in traffic information, but can also be used for navigation. The luminaire can easily transmit position information without the need for a GPS. Then, the system could work as a complement to navigation in dense urban areas (where the GPS has problems) or, for example, for underground parking [13].

However, the system can be also used conversely. If the transmission of the telemetry towards the public lighting luminaire was added, then it would be able to respond to the currently set navigation destination. In case of low night-time operations, the system could dynamically change the output of the public lighting luminaires based on the current requirements. Theoretically, electricity could be used more economically while maintaining the necessary visibility of road users.

Fig. 8. Test of proposed solution (receiver part) of VLC in public lighting in special polygon at the university.

Acknowledgment

The research described in this article could be carried out thanks to the active support of the projects no. SP2020/38 and SP2020/76, VI20172019071. The work has been partially supported by project no. CZ.1.07/2.3.00/20.0217. This work was supported partially by the Ministry of Education, Youth and Sport of the Czech Republic by the project reg. nr. CZ.02.1.01/0.0/0.0/16_019/0000867.

References

[1] L. U. Khan. Visible light communication: applications, architecture, standardization and research challenges. *Digital Communications and Networks*, 3,(2017), pp. 78–88.

[2] A.-M. Cailean, M. Dimian. Current Challenges for Visible Light Communications Usage in Vehicle Applications: A Survey. *IEEE Communications Surveys and Tutorials*, 19(4), (2017), pp. 2681–2703.

[3] L. P. Maguluri, Y. S. V. Sorapalli, L. K. Nakkala and V. Tallari, Smart street lights using IoT, in *Proceeding of 2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, India, Tumkur, 2017, pp. 126–131.

[4] N. Lourenco, D. Terra, N. Kumar, L. N. Alves and Rui L. Aguiar, Visible Light Communication System for outdoor applications, in *Proceeding of 2012 8th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*, Poland, Poznan, pp. 1–6.

[5] D.-R. Kim, S.-H. Yang, H.-S. Kim, Y.-H. Son and S.-K, Han Outdoor Visible Light Communication for inter- vehicle communication using Controller Area Network, in *2012 Fourth International Conference on Communications and Electronics (ICCE)*, Vietnam, Hue, pp. 31–34.

[6] V. M. Baeza, M. Snchez-Fernndez, A. G. Armada and A. Royo, Testbed for a LiFi system integrated in streetlights, in *Proceeding of 2015 European Conference on Networks and Communications (EuCNC)*, France, Paris, pp. 517–521.

[7] F. Knobloch, Noncoherent dimming frequency shift On-Off keying scheme for low data rate optical street lighting communication, in *Proceeding of 2015 17th International Conference on Transparent Optical Networks (ICTON)*, Hungary, Budapest, pp. 1–5.

[8] T. Stratil, P. Koudelka, J. Jankovych, V- Vasinek, R. Martinek, T. Pavelek, Broadband over Visible Light: High power wideband bias-T solution, in *Proceeding of 2016 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, Czech Republic, Prague, pp. 1–5.

[9] T. Stratil, P. Koudelka, R. Martinek and T. Novak. Active Pre-Equalizer for Broadband over Visible Light. *Advances in Electrical and Electronic Engineering*, 15(3), (2017), pp. 553–560.

[10] A. B. Carlson, Communication Systems: *An Introduction to Signals and Noise in Electrical Communication*, 4th ed., McGraw-Hill, 2001.

[11] A. T. Erguzel. A study on the implementation of dimmable street lighting according to vehicle traffic density, *Optik - International Journal for Light and Electron Optics*, 184, (2019), pp. 142–152.

[12] F. Knobloch, Impact of Dimming and Aperture on the Optical Wireless Performance in Public Street Lighting, in *Proceeding of 14th International Conference on Telecommunications - ConTEL 2017*, Croatia, Zagreb, 2017, pp. 27–33.

[13] T. Schaal and E. Zeeb, Optical Free Space Communication with LED Rear Lights, in *Proceeding of 5th International Symposium Progress in Automobile Lighting, Proc. PAL 2003*, Germany, Darmstadt, 2003, pp. 942–954.

3D Visible Light Positioning of an Angle Diverse Receiver based on Track Analysis

Andreas P. Weiss, Franz P. Wenzl
Institute for Surface Technologies and Photonics – Smart Connected Lighting
JOANNEUM RESEARCH Forschungsges.mbH
Industriestraße 6, A-7423 Pinkafeld, Austria
andreas-peter.weiss@joanneum.at
franz-peter.wenzl@joanneum.at

Claude Leiner, Felix Lichtenegger, Christian Sommer
Institute for Surface Technologies and Photonics – Light and Optical Technologies
JOANNEUM RESEARCH Forschungsges.mbH
Franz-Pichler-Straße 30, A-8160 Weiz
claude.leiner@joanneum.at,
felix.lichtenegger@joanneum.at,
christian.sommer@joanneum.at

We present an approach to enable 3D Visible Light Positioning (VLP) utilizing the fingerprinting method combined with the approach of track analysis. Especially in rooms with a highly symmetric arrangement of the luminaires, fingerprinting suffers from high ambiguity of the received signals. We demonstrate an approach to overcome this ambiguity by expanding fingerprinting by track analysis in 3D VLP based on angle diverse receivers. Furthermore, we give a more detailed insight on the impact of the number and the kind of the segmentations of the related optical elements for the receivers and their effect on the performance of the designed algorithm. We demonstrate that by such an approach the success rate for correct 3D position estimation can be enhanced between 29 % and 64 % in dependence of the respective angular segmentation of the receiver.

Keywords— Visible Light Positioning, Angular diverse VLP receivers, 3D positioning

I. INTRODUCTION

Precise Indoor Positioning is the cornerstone of many Location Based Services such as Indoor Navigation and Tracking, Marketing, Emergency Services etc. [1]. For outdoor applications, the Global Positioning System (GPS) has a dominant role, but since GPS based systems cannot facilitate precise positioning data indoors other technologies and methods have been developed for indoor environments in the recent years. Overall, the different systems in this regard can be divided into 4 main categories.

One of the most prominent approaches in the category of Radio-Frequency (RF) based systems are WiFi based ones [2, 3, 4], but also Bluetooth Low Energy (BLE) [5] or Radio-Frequency Identification (RFID) tag [6] based systems (among many more) have been proposed. An in-depth survey on wireless indoor positioning systems and methods is given in [7]. Overall, the performance of RF based systems suffers from signal interference and multipath effects.

The second main category are camera-based systems [8,9], where stationary or mobile cameras are used to determine the position of the camera itself or an object that is inside the Field of View (FoV) of the camera(s). This technology offers high precision in positioning, but has the disadvantage of high computational resource demand and requires high communication bandwidth. Furthermore, this technology raises certain security and privacy concerns.

Approaches utilizing Light Detection and Ranging (LIDAR) [10] and other approaches utilizing for example Ultra Wide band (UWB) [11] sensors or ultrasonic sensors [12] form the third category. They can be summarized as technologies where a pulse is emitted from a transmitter and the reflection of this pulse by an object is then appraised to infer the position of the object.

As an alternative to the above-mentioned approaches, indoor positioning by the means of visible light emitted from the luminaires of the obligatory room lighting has gained a lot of attention in the recent years. This technology is also referred to as Visible Light Positioning (VLP) [13, 14]. VLP has a close relationship with the field of Visible Light Communication (VLC), where by modulating the intensity of the light emitted from a Light emitting Diode (LED) and by reconstruction of this modulation with a photosensitive device (usually a photodiode (PD)) information in the form of data is transferred [15]. With the combination of VLC and VLP it is, besides performing the positioning task itself, also possible to forward the required data for this task, like room size, ID of the luminaire etc. to the receiver. The receiver then performs the task of positioning following one of the three major methods to conduct positioning by VLP.

Geometry based approaches can be divided into the subcategories of signal triangulation or trilateration. By exploiting signal characteristics from multiple light sources, like Received Signal Strength (RSS), Time of Arrival (ToA), Time Difference of Arrival (TDoA) or Angle of Arrival (AoA), the position of the receiver in relation to that of the light sources can be derived. With the absolute position of the light sources known, also the absolute position of the receiver can be deduced.

Proximity based approaches in their most basic implementation are simply detecting the ID sent forth by the light source and by looking for the position of this light source in a stored map, the position can be determined.

The method of Fingerprinting, which is also applied in the context of other wireless data transmission technologies, like WiFi [18], consists of two consecutive phases, an offline

phase and an online one [16, 17]. In the offline phase, a map of the area or of the room, where the position is intended to be detected, is constructed by measuring the values of the incident light on the receiver at a large number of positions. In the online phase (performing the positioning task), the respective measured values of the incident light on the receiver at the respective positions are compared with the values of the offline map and the closest match is considered to be the position of the receiver.

However, such a fingerprinting approach suffers from some certain drawbacks. First, the creation of the offline map is quite time consuming and, second, it is necessary to trade off the resolution of the offline map carefully. On the one hand, the resolution should be high enough to determine the position accurately, but on the other hand, it should be coarse enough so that the points in the offline map do not suffer from high ambiguity. In the following, we propose an approach to overcome the problem of a high ambiguous offline map, by expanding the comparison between the offline map and the received online value sets from only one value set to a concatenated set of values generated by the movement of the receiver. Furthermore, in our approach we do not modulate the light source to send any information between the LED and the receiver.

As mentioned before, ambiguities in the offline map have a very negative effect on the performance of the fingerprinting method with respect to accuracy of positioning. In order to relief this problem, segmented receivers, which can discriminate between the directions from which the light is impinging, have been utilized. Such angle diverse receivers can be designed to achieve such a differentiation by placing the receiving elements on the different sides of shaped objects such as pyramids, cubic shaped objects etc. [19, 20] so that they receive the light from different directions. Another approach to achieve angle diversity is to use an array of photodiodes and an aperture that is placed at some distance above the array to allow for the projection of shadows from the aperture on certain parts of the array, which vary in dependence of the position of the receiver [21]. Furthermore, also systems that utilize optical elements such as spherical lenses have been reported [22].

This paper is divided into the discussion on the preferable angular receiver design for our approach, given in section II. Section III describes the designed algorithm to perform the track analysis. Section IV describes the implemented simulator used for the experimental setup in Section V. In Section VI, the achieved results of our proposed algorithm are presented. This paper is concluded with Section VII on Discussion and Future Works.

II. Receiver Design

In our design approach for an angular diverse receiver, we are utilizing the concept of freeform micro-optical elements (FF-MOEs). Such elements can be designed in ultrathin nature, e.g., to guide the light, as in the present study, towards photosensitive elements. Such FF-MOEs for example can be tailored for the needs in lighting, e.g., to allow for ultra-thin direct-lit luminaires with high distance to height ratios [23 - 25]. The maximal height of suchlike FF-MOEs can be confined to several 10 of micrometers, which enables cost- and time-effective mastering (e.g. grayscale laser lithography) and replication methods like imprinting [26]. With modern high-speed and large-area imprinting technologies, like UV-Nanoimprint-Lithography, a large-scale fabrication of such elements can be envisioned. Here we apply the concept of FF-MOEs for VLP, still we do not focus on their design, we only consider a “segmentation”, this means that light impinging on different areas (segments) of such an optic (in dependence of the angle of incidence) is considered to be forwarded towards different photodiodes. The biggest advantage of such an FF-MOE approach, compared to angular diverse receivers formed by apertures or geometrical shapes, is that the structural form of the whole receiver set-up can be kept comparably small and thin. With such a compact design such receiver set-ups are much easier integrable, e.g., into wristbands worn by persons or even into their cloths. This would make the positioning of persons possible without the need of carrying a device in the hand.

Segmentation of the receiver into certain shapes and forms plays a crucial role in the achievable performance in VLP. The two main approaches for segmentation are the partition of the receiver into segments based on the areas and the angles under which the light impinges on the FF-MOEs. This allows either for a segmentation with respect to the Theta (θ) angles or with respect to the Phi (ϕ) angles, see Fig. 1. Nevertheless, in any case the light impinging on each of the segments has to be forwarded to its associated photodiode and Transimpedance Amplifier (TIA) in order to convert the incoming light impinging on the respective segment into a corresponding electric voltage. Figure 1 shows these two main approaches, with on the left side depicting a Theta based segmentation and on the right side showing a Phi based segmentation.

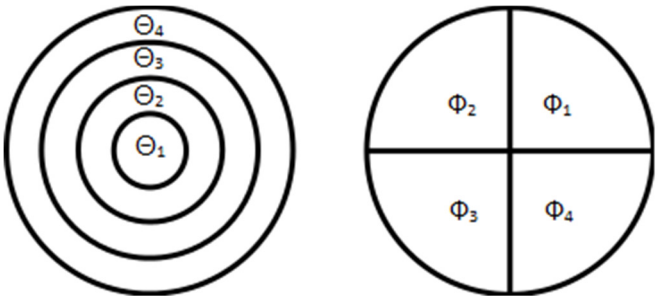


Fig. 1. Left: Receiver optics segmentation with respect to a Theta angle segmentation. Right: Receiver optics segmentation with respect to a Phi angle segementation.

For both designs, we fixed the maximum FoV of the receiver to a theta angle of 60°. The segments are distributed symmetrically. With respect to the exemplary Theta segmentation in Figure 1 (left side), this means that all light rays that impinge on the receiver with the incident angle theta between 0° and 15° will be guided to the PD for segment θ_1 and rays with incident angle theta between 15° and 30° are guided towards the PD of segment θ_2. The PD of segment θ_3 will receive rays with an incident angle between 30° and 45° and finally rays incident with 45° to 60° will be guided towards the PD of segment θ_4. Light rays with a theta angle larger than 60° are considered as lost. In the design shown on the right hand side, Phi segmentation, the incident theta angle has no influence on which PD the rays are guided, except rays with a theta angle > 60°, which are considered as lost too. For Phi segmentations, the area on which the light impinges is evenly distributed among the segments, so every segment has

the same geometrical size. For example, every ray that impinges on the right upper quarter of the receiver area, ϕ_1, will be guided towards the same associated photodiode.

It is intuitive that a high-segmented receiver will be beneficial for VLP positioning since the more segments are present, the more positions of the receiver in a room will be differentiable. In theory, this assumption is true, but there are trade-offs that need to be considered. First, the electronic design of the receiver is more complex and costly the larger the number of segmentations is. Second, when a specific application requires a small overall receiver design and the lateral extension of the receiver optics surface is fixed, the amount of light imping on one segment becomes the less the larger the number of segments becomes. This may pose the problem that in case that the amount of the impinging light is below the detection threshold of the photodiode, no digital output can be generated. Consequently, there is a trade-off between the number of segments and the realizable FF-MOE and electronic designs.

With the two options for angular segmentation, Theta and Phi, and with the pre-requisite that the areas of the segments should not be too small, it is necessary to determine which of the two types of segmentations is more effective in our VLP setting. This was done in a coarse pre-selection study by simulating a room with four symmetrically arranged luminaires (a detailed description of the implemented simulator is given in section IV). The room size has been chosen to be 5 meters by 5 meters in lateral dimensions and 3 m in height, see section V. A very coarse raster of possible receiver positions inside the room with dimensions of 10 cm by 10 cm in x and y positions and 50 cm in z position was simulated for this study on the impact of the type of segmentation. Please note that for the z dimension, the possible receiver positions were only simulated between 0 cm (this means the floor of the room) and 200 cm in height. Above the height of 200 cm and due to the restricted maximum FoV of the receiver, for most of the positions (except directly under or in very close vicinity to a luminaire) in the room no light will impinge on the receiver, and therefore this positions are not usable for VLP. The simulations give a 3-D matrix of light distribution values on the receiver, which has a dimension of 51x51x5. Please note that the values of this matrix, representing receiver positions in the room, are normalized for each position, to rather deal with the relative distribution of the light instead of the absolute numbers. For more details, see section IV. The segmentation of the receiver was set to four Phi segments and four Theta segments, see Figure 1. Since the method of fingerprinting relies on the principle of comparing an actually measured value with a known offline map of values, non-unique value sets of the map pose a problem in this approach. Therefore, the knowledge on which type of angle segmentations gives reason for less non-unique values in this coarse raster pre-selection study, allows to decide which type of segmentation is to be preferred for the set-up of the algorithm described in section III.

The results of the simulations show that for four Theta segments only 3 of the 13005 simulated receiver positions are unique. The 3 unique positions of the Theta segmentation are the exact center position in the room at the heights 0 cm, 50 cm and 100 cm. On the contrast, for four Phi segments, 4872 values are unique. This means for the four Phi segments the investigations show that ~37,5 % of the positions can be correlated unambiguously. Therefore, a Phi based receiver segmentation was chosen for the investigations and discussions in the following. To illustrate the difference in the number of non-unique value-sets, we compared the value-sets for the Theta segmented receiver and the Phi segmented receiver at a height of 100 cm, which constitute matrices having dimensions of 4x51x51. In these matrices, we compared the value-sets for each position to compute how often the same value-set occurs in the matrix. For example, the value-set for position x = 10 and y = 10 for Theta segmentation occurs 3 more times. Figure 2 shows the number of occurrence of the same value-sets for the 51x51 receiver positions at a height of 100 cm assuming Theta segmentation on the left side and Phi segmentation on the right hand side. The colors are in accordance with the respective calculated occurrence numbers of the same value-sets.

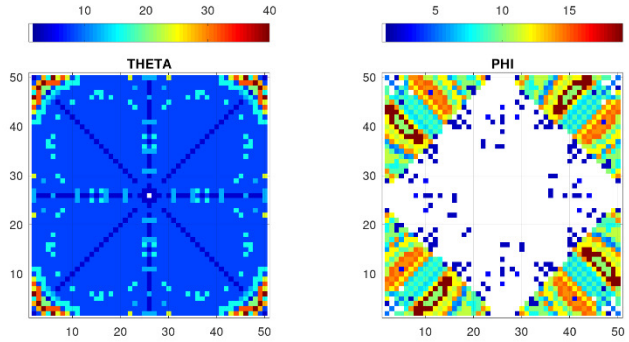


Fig. 2. Left: Number of the occurrence of a position value-set for Theta segmentation at a height of 100 cm. Right: Number of the occurrence of a position value set for Phi segmentation at a height of 100 cm.

In Fig. 2, the positions with a unique value-set are colored white, whilst all other colored points represent positions with non-unique value-sets. With this representation, it is clear to see that the Phi segmented receiver outperforms the Theta segmented receiver in terms of uniqueness.

III. Algorithm Design

Our proposed algorithm is based on the idea that when the position of the receiver is highly ambiguous on the base of only one single value-set, because of a highly symmetric arrangement of the luminaires of the room lighting, this value-set has to be expanded. As it was shown in the previous section, the amount of the light impinging on the receiver is not unique for many positions in the room. In the fingerprinting method, this would lead to highly ambiguous position determination since the “measured” values can be found at various positions of the offline map. To overcome this problem we propose an algorithm that is not solely relying on a single measurement, but on consecutive readings due to the movement of the receiver in the room. This gives reason for an enlarged number of value-sets. This series of value-sets is then compared with those of hypothetic movements, by calculating the Euclidian distance between the value-sets.

The algorithm consists of two phases, starting with the construction of the offline map. This offline map holds the values for every Phi segment at certain positions in the room. The next phase is the online phase, which consists of four steps. First, the values for every Phi segment of the receiver are acquired. This value-set is then compared to the offline map to find possible positions of the receiver in the room.

Every position in the offline map that matches the current measured values is marked as a possible track starting point. In the second step, the possible tracks with the origin in these possible starting points are computed. So for example if two points in the offline map are marked as possible starting points, tracks originating from these two points going left, right, up, down etc. are computed. The tracks will be stored in a matrix, further on called the hypotheses matrix. This matrix holds for every starting point the hypotheses with the defined number of track steps. For example, if the receiver has four Phi segments and the number of track steps is defined as 5 and with the above mentioned example of two possible starting points and six track hypotheses, in total the hypotheses matrix will have the dimensions of 4x5x6x2. In the third step, when the receiver is moving, the current data from the receiver are acquired. With the above-mentioned example of four Phi segments this results in a 4x1 vector. This vector is now concatenated with the value-set of the first stage (origin of the movement) to form a 4x2 matrix. With the next step of the receiver, another 4x1 vector is acquired and concatenated to the matrix, resulting in a 4x3 matrix, etc. This third step is repeated until the defined number of track steps is reached. In our example with the number of track steps set to 5, it will be repeated until a 4x5 matrix has been created. In the fourth and final step, the Euclidian distance between the measured matrix and the hypothesis for every track originating from every possible starting point is computed. To reuse the example from above again, the hypotheses matrix has the dimensions of 4x5x6x2, with the third dimension giving the number of hypotheses and the fourth dimension giving the starting points. The acquired measured data have the dimensions of 4x5. So for every hypothesis from every start point, the Euclidian distance between the hypothesis and the measured data is calculated. The hypothesis that yields the minimal Euclidian distance to the actual values is then giving the actual starting point and the most probable movement of the receiver in the room.

The algorithm was implemented in GNU/Octave and works with six different directions for the track hypotheses buildup. The directions of the possible tracks are going into x direction to the right, going into x direction to the left, moving forward in the y direction and moving backward in the y direction. For these hypotheses, no movement in height is assumed. The fifth and sixth possible movements are going straight up and going straight down, with no movement in the x or y directions. If a possible track goes out of the sizes of the room in either one of the three dimensions, the track is dismissed. The proposed algorithm is tested with simulated data created with the implemented Simulator described in the next section.

IV. Simulator

For this work, a simulator has been implemented in Matlab that can simulate an environment in which our approach is applied. The simulator can be adjusted freely to different room sizes and positions of the luminaires. The output power of the light sources in Lumen can also be chosen freely. The adjustable parameters of x resolution, y resolution and z resolution form a raster inside the defined room. For every point of this raster the incident light on the receiver plane is computed based on an assumed Lambertian emission pattern of the light sources. The angle of radiation of the light sources is fixed with 120°. In this simulation environment, the horizontal orientation of the receiver is assumed straight upward, with no tilt. The incident light on the receiver is then split into the individual segments. The number of segments of the receiver can also be defined arbitrarily. In order to avoid the use of absolute numbers for the amount of light impinging on the segments, we chose to normalize the values for each receiver position. With this approach, we do not depend on absolute values in our algorithm, since any change in the output power of the luminaires would change these absolute values and would make a complete new simulation run necessary. The output is the distribution of the total impinging light on the individual segments for each possible receiver position in the room, according to the defined raster. For example if at a certain receiver position the incident light has the total power of 4 Watt, which distributes among the exemplary 4 Phi segments with ϕ_1=0.8 W, ϕ_2=0.6 W, ϕ_3=1.2 W and ϕ_4=1.4 W, the normalized output of this position will be ϕ_1= 0.20, ϕ_2=0.15, ϕ_3=0.30 and ϕ_4=0.35. These values, which are rounded to the second digit after the comma, are stored in a matrix with the size according to the defined room size and the chosen raster of the receiver positions.

Reflections of the light on the walls, the floor or the ceiling are not considered in this simulation environment. Furthermore, no modulation of the light emitted from the luminaire is considered.

V. Experimental Setup with Simulated Data

For the experimental verification of our approach, an exemplary application in a room with a highly symmetric arrangement of the luminaires is assumed. The following parameters for the room were chosen for the simulations:

- The dimensions of the room are 5 meters in width, 5 meters in length and 3 meters in height
- The room lighting consists of 4 Luminaires with a total output power of 4000 Lumen
- The positions of the luminaires are symmetrical with respect to the center of the ceiling plane
- Each luminaire has a size of 12 cm by 12 cm
- The resolution of the possible receiver positions is 5 cm both in x and y directions
- The height resolution is 10 cm ranging from the height of 0 (with respect to the room floor) to 2 meters in the z axis.

Figure 3 shows a model of the room with the Luminaires named L1 to L4.

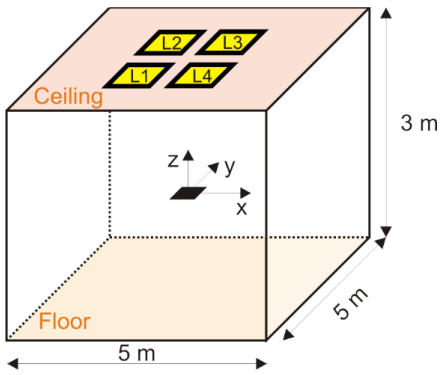


Fig. 3. Sketch of the simulation model of the room

For the determination of the effects of a different number of Phi segments of the FF-MOE and different number of track steps on the performance of the proposed algorithm, we performed the experiments for Phi segmentations ranging from four segments to a maximum of eight segments. For each of the different segmentations the whole track ranges from 1 to a maximum of 20 steps along the raster. With the defined raster of (5cm x 5cm x 10cm) this results in a maximum movement of 1 meter in x or y directions and 2 meters in z direction. Each of the scenarios, e.g. a segmentation with 5 Phi segments and the number of track steps = 7, was performed for 1000 times.

For each run, the real position of the receiver was chosen randomly inside the raster of the room. Then the algorithm was performed as described in section III. First, the possible start points that match the values at the random real position were sought after in the offline matrix. For each of these points the movement hypotheses were built. In the next step, the real movement of the receiver was chosen randomly out of the defined 6 movements. If the randomly chosen movement is not possible, since it would go out of the sizes of the simulated room, it was neglected and another movement was chosen. Afterwards the Euclidian distances for each step between the value sets from the "real" movement and the track hypotheses was computed. The track with the lowest sum of computed distances is considered as the result. If two tracks have the same lowest value, the result was marked as ambiguous and consequently counted as a wrong estimation. In the final step, the determined result was compared to the "real" starting point of the movement. If the two points match, the result is considered as correct. Figure 4 shows the achieved percentage of correct determinations of the starting point on the y-axis. The different number of track steps are given on the x-axis. The curves represent the Phi segmentations with the labels of e.g. Phi 4 meaning that the receiver is has 4 segments; Phi 5 having 5 segments and so forth.

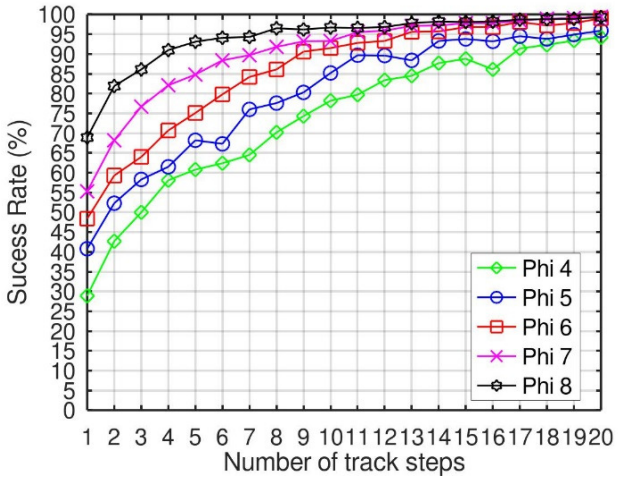


Fig. 4. Percentage of correct determinations of the starting points for different Phi segmentations in dependence of the number of track steps

VI. Results

The results in Figure 4 show the validity of our approach, since it is clear to see that for any receiver design the percentage of the correct determinations increases with an increased number of track steps. This circumstance can be explained by the simple fact that the longer the track becomes, the higher becomes the probability that combined value-sets of the track are unique. With 20 steps for the track, which would correlate to a movement in the x and y directions of 1 meter in real world or a movement of 2 meters in the z axis, all investigated number of Phi segmentations achieve a correct determination rate of over 94 % with the maximum of Phi segmentations, Phi 8, reaching over 99 %. The differences regarding the starting points of the different curves (number of track steps = 1) is based on the fact that the more segments a receiver holds, the more unambiguous are the values compared to the possible receiver positions in the room. This gives the advantage that for even smallest numbers of track steps (5cm in x or y or 10 cm in z direction) the accomplished success rate varies between ~ 30 % for 4 Phi segments and ~ 70 % for 8 Phi segments. This behavior is not surprising, still the results of our investigation allow to get an idea on the amount of improvement of correct determinations with increasing number of segments.

For the further design of the receiver the results from this work allow to find a careful trade-off between the more complex electronic design and the higher costs for a segmented receiver with a larger number of segments and the computational complexity of the algorithm that increases with a larger number of track steps. From the viewpoint of the FF-MOEs, a higher segmentation would not necessarily cause higher costs in production. Contrarily, for the electronic design the number of necessary components and therefore the cost increases with every segment, since each segment requires a dedicated PD and TIA amplifier. The increase in computational complexity is based on the fact, that, the larger the chosen number of track steps, the larger becomes the second dimension of the created hypotheses matrix, resulting in an increased amount of data to be stored for the hypotheses matrix and consequently the computational effort for the calculation of the Euclidian distance is increasing. Smaller numbers of track steps will also be beneficial in an aspired real world implementation, since the shorter the track, the more likely it is that the real movement of the receiver corresponds with one of our 6 movement hypotheses.

Overall, it can be summarized that our approach of utilizing track analysis combined with fingerprinting in an exemplary room (5m x 5m x 3m) with a resolution of 5cm x 5cm x 10 cm gives reason for an improvement of correct determinations of the receiver starting positions from ~30 % up to 94 %, for receiver designs with 4 Phi segments. This means that applying solely fingerprinting allows to infer a randomly chosen origin position of the receiver in our model room only for ~30% of possible positions. In combination with the approach discussed in this study, 94 % of the random origin positions can be determined correctly. Even for receivers with a larger number of segments, e.g., 8 segments, the improvement of correct determinations is ~ 29 %.

For the results of our application, it is possible to define a threshold of probability in origin point estimation and then find possible receiver segmentations and a minimal number of track steps. For example, consider that a minimum probability for the estimation of correct starting points of 90% is chosen. This leads to the following numbers of receiver segmentations in combination with the minimal number of required track steps: Phi 4 – Number of track steps 17, Phi 5 – Number of track steps 14, Phi 6 – Number of track steps 9, Phi 7 – Number of track steps 8 and Phi 8 – Number of track steps 4.

VII. Discussion and Future Works

In this study, we presented a 3D VLP positioning method based on the fingerprinting in combination with track analysis, utilizing an angle diverse receiver based on FF-MOEs. With our simulation-based results, we were able to demonstrate the validity of our approach and to show how different (numbers of) segmentations of an angle diverse receiver influence the achievable results.

The main objective of this work is on the correct determination of the starting position of the receiver. In the simulated data, reflections from the wall or the ceiling of the room have not been considered, nor obstacles blocking the line of sight. It is undoubtable that in a communication scenario the signal multipath propagation will have a negative effect on the communication capabilities. With the task of positioning based on fingerprinting in focus, such effects do not necessarily have a negative effect, they could be even helpful to overcome the problem of non-unique values in the offline map, since the main causes for unambiguity are the highly symmetrical arrangements of the luminaires and the rooms in the chosen set-up.

In future works we plan to expand the simulation tool, so that we will be able to simulate also the related electronic performance of the receivers, e.g., including noise. With this expansion, the system and the algorithm can be further developed towards real world implementation. Additionally, we will investigate the incorporation of IMU sensors into the set-up in order to provide movement data from the IMU sensor to the algorithm of track hypotheses generation. For example if the IMU sensors acquire that the receiver is moving into direction north, the algorithm can dismiss all the hypotheses that have been built up going into a different direction. This will reduce the computational load on the receiver unit and therefore will allocate resources for a more detailed track hypotheses. After all these enhancements, we will fabricate the whole set-up in order to demonstrate its applicability and to compare its experimental performance with the simulated predictions.

Acknowledgment

Project “SmartLight2Live” is part-financed by the Federal State of Austria, the Province of Burgenland and the European Regional Development Fund in the context of the Investment for growth and jobs goal.

References

[1] A. Basiri, E. S. Lohan, T. Moore, A. Winstanley, P. Peltola, C. Hill, A. Pouria and P. Silva, Indoor location based services challenges, requirements and usability of current solutions. Computer Science Review, vol 24, 2017, doi:10.1016/j.cosrev.2017.03.002.

[2] C. Yang and H. Shao, "WiFi-based indoor positioning," in *IEEE Communications Magazine*, vol. 53, no. 3, March 2015, pp. 150-157.

[3] G. Caso, L. De Nardis, F. Lemic, V. Handziski, A. Wolisz and M. D. Benedetto, "ViFi: Virtual Fingerprinting WiFi-Based Indoor Positioning via Multi-Wall Multi-Floor Propagation Model," in *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1478-1491, 1 June 2020, doi: 10.1109/TMC.2019.2908865.

[4] C. Own, J. Hou and W. Tao, "Signal Fuse Learning Method With Dual Bands WiFi Signal Measurements in Indoor Positioning," in *IEEE Access*, vol. 7, 2019, pp. 131805-131817

[5] A. Filippoupolitis, W. Oli and G. Loukas,”Occupancy detection for building emergency management using BLE beacons”, International Symposium onComputer and Information Sciences, 2016, pp. 233-240,2016

[6] N. Li, G. Calis and B. Becerik-Gerber,”Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations”, Automation in Construction 24, 2012, pp. 89-99

[7] F. Zafari, A. Gkelias and K. K. Leung, "A Survey of Indoor Localization Systems and Technologies," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, thirdquarter 2019. pp. 2568-2599

[8] Y. Moon et al., "CaPSuLe: A camera-based positioning system using learning," 2016 29th IEEE International System-on-Chip Conference (SOCC), Seattle, WA, 2016, pp. 235-240.

[9] H. Deng, Q. Fu, Q. Quan, K. Yang and K. Cai, "Indoor Multi-Camera-Based Testbed for 3-D Tracking and Control of UAVs," in *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 6, pp. 3139-3156, June 2020, doi: 10.1109/TIM.2019.2928615.

[10] Y. T. Wang, C. C. Peng, A. Ravankar, A. Ravankar, “A Single LiDAR-Based Feature Fusion Indoor Localization Algorithm”. In Sensors. vol 18., 2018, pp. 1294, doi:10.3390/s18041294

[11] S. N. A. Ahmed and Y. Zeng, "UWB positioning accuracy and enhancements," *TENCON 2017 - 2017 IEEE Region 10 Conference*, Penang, 2017, pp. 634-638.

[12] A.D. Lazarov, D. Minchev, A. Dimitrov, ”Ultrasonic Positioning System Implementation and Dynamic 3D Visualization”. Cybernetics and Information Technologies., vol. 17, 2017, doi:10.1515/cait-2017-0023

[13] T. H. Do, M. Yoo, “An in-Depth Survey of Visible Light Communication Based Positioning Systems”, in Sensors, vol. 16, 2016, pp. 678. doi:10.3390/s16050678

[14] Y. Zhuang et al., "A Survey of Positioning Systems Using Visible LED Lights," in IEEE Communications Surveys & Tutorials, vol. 20, no. 3, thirdquarter 2018, pp. 1963-1988, doi: 10.1109/COMST.2018.2806558

[15] L. E. M. Matheus, A. B. Vieira, L. F. M. Vieira, M. A. M. Vieira and O. Gnawali, "Visible Light Communication: Concepts, Applications and Challenges," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, 2019, pp. 3204-3237

[16] T. Wenge, M. T. Chew, F. Alam and G. S. Gupta, "Implementation of a visible light based indoor localization system," *2018 IEEE Sensors Applications Symposium (SAS)*, Seoul, 2018, pp. 1-6.

[17] Y. Chen, W. Guan, J. Li and H. Song, "Indoor Real-Time 3-D Visible Light Positioning System Using Fingerprinting and Extreme Learning Machine," in *IEEE Access*, vol. 8, 2020, pp. 13875-13886

[18] W. K. Zegeye, S. B. Amsalu, Y. Astatke and F. Moazzami, "WiFi RSS fingerprinting indoor localization for mobile devices," *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, 2016, pp. 1-6.

[19] B. Xie, G. Tan, Y. Liu, M. Lu, K. Chen, T. He, “LIPS: A Light Intensity--Based Positioning System for Indoor Environments”. ACM Transactions on Sensor Networks, vol. 12, 2016, doi:10.1145/2953880

[20] J. Vongkulbhisal, B. Chantaramolee, Y. Zhao, W. Mohammed, “A fingerprinting ‐ based indoor localization system using intensity modulation of light emitting diodes”, Microwave and Optical Technology Letters, vol. 54, 2012, doi:10.1002/mop.26763

[21] S. Cincotta, C. He, A. Neild, J. Armstrong, “Indoor Visible Light Positioning: Overcoming the Practical Limitations of the Quadrant Angular Diversity Aperture Receiver (QADA) by Using the Two-Stage QADA-Plus Receiver”, in Sensors, vol. 19, 2019, doi:10.3390/s19040956

[22] T. Q. Wang, Y. A. Sekercioglu and J. Armstrong, "Analysis of an Optical Wireless Receiver Using a Hemispherical Lens With Application in MIMO Visible Light Communications," in *Journal of Lightwave Technology*, vol. 31, no. 11, 2013, pp. 1744-1754

[23] C. Leiner, W. Nemitz, S. Schweitzer, F. P. Wenzl, L. Kuna, F. Reil, P. Hartmann, C. Sommer, Thin direct-lit application for general lighting realized by freeform micro-optical elements, Proc. of SPIE 9955, 99550E, (2016)

[24] C. Leiner, W. Nemitz, S. Schweitzer, F.-P. Wenzl, C. Sommer, Smart freeform optics solution for an extremely thin direct-lit backlight application**,** Proc. of SPIE 9889, 988911, (2016)

[25] C. Leiner, W. Nemitz, F. P. Wenzl, C. Sommer, Ultrathin free-form micro-optical elements for direct-lit applications with a large distance-height ratio, OSA Continuum 1, 1144-1157 (2018)

[26] L. Kuna; C. Leiner; W. Nemitz; F. Reil; P. Hartmann; F.P. Wenzl; C. Sommer, Optical design of freeform micro-optical elements and their fabrication combining maskless laser direct write lithography and replication by imprinting, J. Photon. Energy. 7, 016002, (2017)

Modelling the Refractive Index Structure Parameter: A ResNet Approach

Christopher Lamprecht, Pasha Bekhrad, Hristo Ivanov, Erich Leitgeb

Graz University of Technology, Institute of Microwave and Photonic Engineering

Inffeldgasse 12, 8010 Graz, Austria

Email: christopher.lamprecht@tugraz.at, bekhrad@tugraz.at, hristo.ivanov@tugraz.at, erich.leitgeb@tugraz.at

***Abstract*—Various atmospheric effects have a negative influence on optical signals, especially in the troposphere, which must be taken into account in free space optical (FSO) communication systems. To obtain a quantitative estimate of these effects, different mathematical models are used, often based on empirical data from around the world. The main problem with existing models is the limited accuracy, due to the different meteorological conditions at different locations on earth. We propose a new approach of modelling the refractive index structure parameter using residual neural networks (ResNets). New models, tailored to the meteorological conditions at any place on earth, can be easily created, which yields in a more accurate estimation of the refractive index profile.**

***Index Terms*—Artificial neural networks, atmospheric turbulence, refractive index structure parameter, ResNet, machine learning, Hufnagel-Valley model**

I. Introduction

Due to the urge of achieving higher data rates, the importance of free space optical (FSO) communication has increased in the last few years. Especially in aerospace applications [1] [2], where large amounts of data need to be transmitted in a timely manner. In optical earth-space communication scenarios [3], the light propagates through the atmosphere inducing additional losses due to fog, clouds or precipitation events [4]. Atmospheric properties change over time due to a change of temperature, pressure, wind, humidity and dew point. These changes reflect on the signal quality of the FSO system, which leads to fluctuations of the signal power at the receiver side. Especially in optical communication systems, fluctuations of the refractive index within the troposphere can lead to atmospheric turbulence effects like beam wandering [5] or scintillation [6], resulting in non-negligible losses on the optical communication channel. For a FSO system design it is therefore necessary to take atmospheric turbulences [7] into account. This is usually done using mathematical models like the Hufnagel-Valley model [8], which accuracy is often limited. To address this problem we propose a new approach based on residual neural networks (ResNets) [9]. The main advantage of neural networks is their ability to find correlations between different data sets and to create hypotheses out of these correlations. A new turbulence model is created by training a neural network using the refractive index structure parameter C_n^2 as target data in relation to the height above sea level and the corresponding wind speed. Since the height and the wind speed correlate with the refractive index structure parameter C_n^2, a hypotheses for C_n^2 can be found eventually. Hence, after the training process the new turbulence model is able to predict the refractive index structure parameter C_n^2 from the height and the wind speed at a given time. If meteorological data is available for different altitudes, the refractive index profile of the troposphere can be recreated using the proposed ResNet model.

II. Atmospheric Effects

The intensity of optical turbulence is measured by refractive index structure parameter C_n^2. Different locations can have different temperature distributions, which affects the C_n^2 turbulence profile. Below the tropopause, the largest gradient of temperature, together with the largest values of atmospheric pressure are considered close to ground, thus at sea level it should expect the largest values of C_n^2. In other words, an increasing altitude results in a decrease of the temperature gradient and thus also a decrease of the values of C_n^2. Taking into account the temperature dynamic during the day, turbulence should be stronger around noon. Due to thermal equilibrium at sunset and dawn, lower values of C_n^2 are expected. Typical values of C_n^2 at ground level can be as low as $10^{-17}m^{-2/3}$ for weak turbulences, whilst strong turbulences can lead to values up to $10^{-13}m^{-2/3}$.

The Hufnagel-Valley model is a heuristic model of the C_n^2 profile to describe the average near-ground turbulence conditions. For this turbulence approximation the Bufton model [8] is sometimes used to model the root mean square of the wind profile v_{rms}, since the wind speeds at varying heights are usually unknown. However, our proposal is using wind speeds obtained from atmospheric RAwinsonde OBservation (RAOB) measurements. Due to its simplicity and the low amount of input parameters, the Hufnagel-Valley model is commonly used as a first estimate of the turbulence profile in earth-space communication scenarios. The Hufnagel-Valley model (stated in [10]), used for data generation and validation of the ResNet model, is given as follows where the last exponential term

accounts for the boundary layer correction:

$$C_n^2(h) = 0.00594 \left(\frac{v_{wind}}{27}\right)^2 \left(10^{-5} \cdot h\right)^{10} e^{-\frac{h}{1000}} + 2.7 \cdot 10^{-16} \cdot e^{-\frac{h}{1500}} + A \cdot e^{-\frac{h}{100}} \quad (1)$$

Using v_{wind} as the measured wind speed at a certain altitude in $[m/s]$ and h is the height above sea level in m. A denotes the turbulence strength at the ground level and is set to $A = 1.7 \cdot 10^{-14} m^{-2/3}$.

The Hufnagel-Valley model applies to the whole atmosphere, although a good description of near ground turbulence is not provided by the boundary layer extension (last term in (1)). Applications which depend heavily on an accurate turbulence modelling in the lowest layer of the atmosphere, should not use the Hufnagel-Valley model. In addition, mid-latitude regions are better suited for the model.

III. Model description

A. ResNet Idea

Deep artificial neural networks are harder to train compared to shallow networks, due to the vanishing/exploding gradient problem [11] [12]. Another issue with deeper networks is the degradation problem, where the accuracy of the model decreases with an increasing network depth [13] [14]. A mitigation to these problems is the ResNet approach. A ResNet building block, as proposed in [9], is shown in Fig. 1 consists of two weight matrices $\mathbf{w_1}$ and $\mathbf{w_2}$ with the corresponding bias vectors $\mathbf{b_1}$ and $\mathbf{b_2}$. The input vector is denoted as $\mathbf{x}$ and the output vector is $\mathbf{y}$. A rectified linear unit (ReLU) is used as an activation function, which is given as follows:

$$ReLU(x) = \max(0, x) \quad (2)$$

The ResNet building block can also be expressed as follows:

$$\mathbf{y} = ReLU\left(\mathbf{w_2} ReLU\left(\mathbf{w_1 x} + \mathbf{b_1}\right) + \mathbf{b_2} + \mathbf{x}\right) \quad (3)$$

The main difference of residual networks to regular feed forward networks is the shortcut connection also denoted as *identity mapping*. In regular deep feed forward neural networks the error gradient vanishes during back propagation, resulting in marginal weight updates in the first few layers. Due to this identity mapping in ResNets, the gradient is preserved during back propagation. As a result, deep networks can be trained with satisfying results, which perform at least as good as shallow networks without degradation of performance.

B. Model Architecture

The ResNet model shown in Fig. 2 consists of a dense layer with two neurons using *tanh* activation functions. The resulting output is then fed into the ResNet structure, which consists of ten sequential ResNet building blocks. Each ResNet building block includes 50 hidden neurons at each weight layer. The output of the ResNet structure is fed into the linear output neuron, which delivers the corresponding C_n^2 value to a given input. The model expects the height above sea level h and the wind speed v_{wind}, at this specific altitude, in order to compute the corresponding refractive index structure parameter C_n^2.

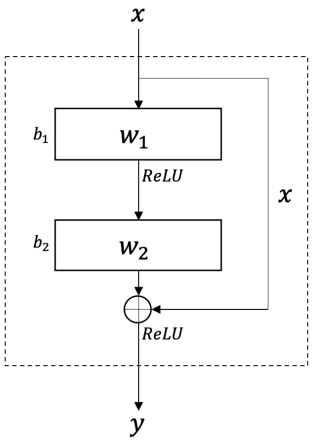


Fig. 1. ResNet building block

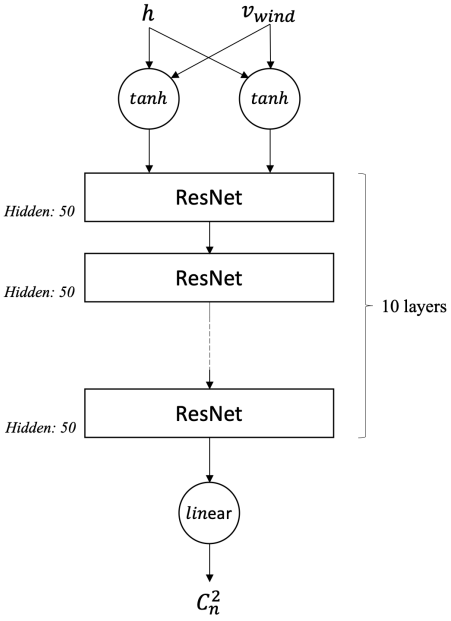


Fig. 2. ResNet model architecture used for modelling the refractive index profile

One advantage of using neural networks for refractive index structure parameter modelling is the independence of input parameters. The input parameters can be easily extended with other parameters like temperature, pressure, humidity or dew point, which might yield in better results while using the same training process and network architecture. Since the ResNet model is trained using data which correlates with the refractive index structure parameter, the performance of the model depends on the quality and amount of data samples.

The Keras API was used for the implementation of the ResNet model in Python. Due to the clear and precise model definition using the Keras API, the implementation effort of the model is quite low.

C. Training and Validation Set

The training and validation target data is generated from the Hufnagel-Valley model using wind speeds v_{wind} of atmospheric RAOB measurements [15] from different observatory locations in order to achieve a better generalization of the final ResNet model. Meteorological data from observatories in Oberschleissheim-Munich (DEU), Güímar-Tenerife (ESP), Hilo-Hawaii (USA) and Wagga Wagga-New South Wales (AUS) from 1^{st} January 2018 to 31^{st} December 2019 are being used. One data sample consists of a height value h and a wind speed value v_{wind}, which represents the input. The corresponding C_n^2 value represents the target value used for the training process. In total 528438 data samples are used for training, which are shuffled and split up into a training set and a validation set of 422750 and 105688 data samples respectively. It is believed that more data for training and validation leads to better generalization of the model. In order to increase the learning performance of the network, the input is normalized to zero mean and unit variance. This normalization of the input can significantly accelerate the training process due to the gained symmetry of the cost function with respect to the weights. Hence, larger training steps can be performed, which leads to larger weight updates and to a faster convergence of the training and validation error.

IV. Results

A. ResNet training results

The proposed ResNet model is trained for 50 epochs using a mini batch size of 50 samples. As a loss function the mean squared error is used, which is minimized using the Adam optimizer [16]. The Adam optimizer only requires first-order gradients for the weight updates and computes individual adaptive learning rates for first and second-order moments, which leads to a better training performance and generally faster convergence of the training error compared to other optimizers like stochastic gradient descent. The final mean squared error (MSE) and the mean absolute error (MAE) of the training and validation set can be seen in Table I. In Fig. 3 a stable training process can be observed, as the training and validation error are steadily decreasing after every training iteration. With a final training mean squared error of $3.2787 \cdot 10^{-6}$ and a validation mean squared error of $5.4738 \cdot 10^{-7}$, the predicted C_n^2 is expected to be close to the real C_n^2, which also the model results show. Although many training and validation samples have been used, the training effort is very low due to the rapid convergence of the training and validation error. One ResNet model can be trained within 10 to 20 minutes.

TABLE I
Model training results after training

	MSE	***MAE***
Training error	$3.2787 \cdot 10^{-6}$	$5.3860 \cdot 10^{-4}$
Validation error	$5.4738 \cdot 10^{-7}$	$4.6168 \cdot 10^{-4}$

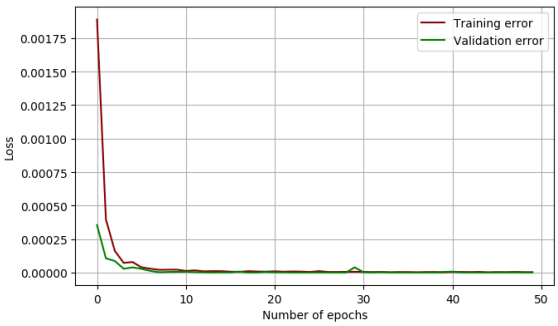


Fig. 3. Training and validation error (MSE)

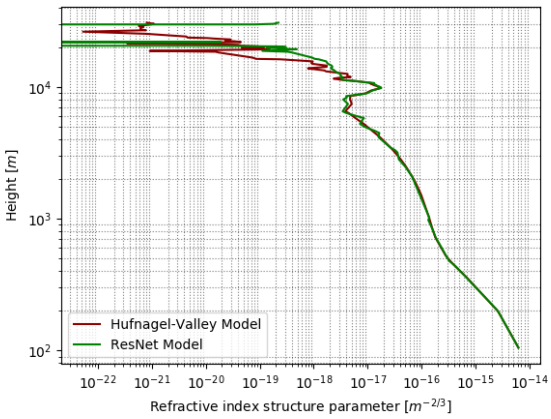


Fig. 4. Hufnagel-Valley model vs. ResNet model using meteorological data from Güímar-Tenerife (ESP) captured on 10^{th} March 2020 averaged over 12 hours from 12:00-24:00 o'clock.

B. Model performance

An advantage of neural networks is the ability to find correlations between different data sets, which is especially helpful in regression tasks like this one. In Fig. 4 the Hufnagel-Valley model is compared to the trained ResNet model, using meteorological RAOB measurements from Güímar-Tenerife (ESP). It can be seen that the ResNet model behaves very similar to the Hufnagel-Valley model. However, above 14000 m the ResNet model differs from the Hufnagel-Valley model. One reason for this difference might be the lack of meteorological data at heights above 14000 m, which leads to inaccuracies during the training of the ResNet model. Since atmospheric models are mostly used to estimate the refractive index profile of the troposphere, which is below 12000 m, this divergence is not considered as a big issue in most applications. Güímar-Tenerife (ESP) is located at 289 m above sea level, which is far beneath the capping inversion layer characterized with around 1 km elevation. When the specifics of Tenerife coastal climate are taken into account, this leads to conclusion that an FSO communication link is highly vulnerable to observed clouds, which can easily block the link. Consequently, a good location for an optical ground station (OGS) is Teide Observatory at 2400 m above sea level, which is extensively used for deep space communication campaigns. At those high

elevations, only atmospheric turbulence and cirrus clouds, which are characterized with an average optical attenuation are observed. Nevertheless, the utilized RAOB measurements and the currently accomplished analysis are completely valid for the entire island due to covering all troposphere including the heights above 2400 m at which the OGS is located.

V. Conclusion

In earth-space FSO communication systems atmospheric turbulences can induce high losses, which need to be considered in the design of the FSO system itself. Therefore, the prediction of the refractive index structure parameter C_n^2 is important to estimate the effects of atmospheric turbulences. Although various mathematical models already exist, we have shown that a machine learning approach for refractive index structure parameter modelling can deliver acceptable results. The bottleneck to good model performance is the training and validation data, which should ideally be collected from many different places on earth in order to deliver accurate results.

References

[1] D. M. Cornwell, "Nasa's optical communications program for 2017 and beyond," in *2017 IEEE International Conference on Space Optical Systems and Applications (ICSOS)*, 2017, pp. 10–14.

[2] S. Yamakawa, Y. Chishiki, Y. Sasaki, Y. Miyamoto, and H. Kohata, "Jaxa's optical data relay satellite programme," in *2015 IEEE International Conference on Space Optical Systems and Applications (ICSOS)*, 2015, pp. 1–3.

[3] E. Samain, D. Phung, N. Maurice, D. Albanesse, H. Mariey, M. Aimar, G. M. Lagarde, N. Vedrenne, M. Velluet, G. Artaud, J. Issler, M. Toyoshima, M. Akioka, D. Kolev, Y. Munemasa, H. Takenaka, and N. Iwakiri, "First free space optical communication in europe between sota and meo optical ground station," in *2015 IEEE International Conference on Space Optical Systems and Applications (ICSOS)*, 2015, pp. 1–7.

[4] E. Leitgeb, M. Gebhart, P. Fasser, J. Bregenzer, and J. Tanczos, "Impact of atmospheric effects in free space optics transmission systems," *Proc SPIE*, 04 2003.

[5] L. Hudcova and P. Barcik, "Experimental measurement of beam wander in the turbulent atmospheric transmission media," in *Proceedings of 22nd International Conference Radioelektronika 2012*, 2012, pp. 1–4.

[6] Z. Nazari, A. Gholami, Z. Vali, M. Sedghi, and Z. Ghassemlooy, "Experimental investigation of scintillation effect on FSO channel," in *2016 24th Iranian Conference on Electrical Engineering (ICEE)*, 2016, pp. 1629–1633.

[7] N. Xiaolong, Y. Haifeng, L. Zhi, C. Chunyi, M. Ce, and Z. Jiaxu, "Experimental study of the atmospheric turbulence influence on FSO communication system," in *2018 Asia Communications and Photonics Conference (ACP)*, 2018, pp. 1–3.

[8] L. Andrews and R. Phillips, *Laser Beam Propagation Through Random Media*, 01 2005.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[10] M. Uysal, C. Capsoni, A. Boucouvalas, and E. Udvary, *Optical Wireless Communications An Emerging Technology*, 01 2017.

[11] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 249–256, 01 2010.

[13] K. He and J. Sun, "Convolutional neural networks at constrained time cost," 06 2015, pp. 5353–5360.

[14] R. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 05 2015.

[15] University of Wyoming, "RAwinsonde OBservation measurement data base," [online], Available: http://weather.uwyo.edu/upperair/sounding.html, [Accessed May 8, 2020].

[16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

A Broadband 2.1 GHz LDMOS Power Amplifier with 700 MHz Bandwidth Implementing Band-pass Filter-Based Matching Networks

*Note: Sub-titles are not captured in Xplore and should not be used

1st Jose Romero Lopera
Institute of microwave and Photonics Engineering (IHF)
Graz university of Technology (TU Graz)
Graz, Austria
romerolopera@tugraz.at

2nd Michael Ernst Gadringer
Institute of microwave and Photonics Engineering (IHF)
Graz university of Technology (TU Graz)
Graz, Austria
michael.gadringer@tugraz.at

3rd Erich Leitgeb
Institute of microwave and Photonics Engineering (IHF)
Graz university of Technology (TU Graz)
Graz, Austria
erich.leitgeb@tugraz.at

4th Wolfgang Bösch
Institute of microwave and Photonics Engineering (IHF)
Graz university of Technology (TU Graz)
Graz, Austria
wbosch@tugraz.at

***Abstract*—In this paper we are presenting a power amplifier based on an RF power LDMOS transistor at 2.1 GHz with a bandwidth of 700 MHz implementing broadband impedance matching networks (BIMN) in band-pass form. The device parasitics modelled as a series RLC structure are absorbed in the input and output matching networks by proper design of matching networks in band-pass form using two different synthesis methodologies. The S_{21} of the amplifier shows 11.2 dB of maximum small signal gain with 2.4 dB ripple along a 700 MHz bandwidth at 2.1 GHz and S_{11} better than -7 dB for 1.64-2.34 GHz. The band-pass matching networks are entirely designed by transmission lines modeling the behavior of the resonant parallel shunt and series LC structures that characterize a band-pass circuit. The amplifier should be able to deliver 1W of maximum output power at saturation.**

***Index Terms*—Broadband impedance matching network, LDMOS, Band-pass, Power amplifier.**

I. Introduction

Modern wireless communication systems are under an ever-increasing demand of bandwidth requirements. Transmission of higher data-rates, operation at different neighbored frequency bands or co-existence of multiple communication protocols are all scenarios that require of higher bandwidth requirements in wireless communication links. The power amplifier (PA) is one of the most critical elements of such communication links, having the highest consumption of DC power, therefore high efficiency is usually desired. However, the simultaneous optimized performance in terms of gain, power, efficiency and bandwidth is not achievable, and some level of tradeoff is usually required, especially when higher bandwidths are desired.

Broadband matching of PA has been a major focus of interest for many years making use of the existing broadband matching theories [1], [2]. Several studies [3], have implemented solutions for broadband matching of PA making use of low pass ladders for the input and output matching networks, which appears to be a reasonable solution to absorb the shunt capacitance and series inductance related to the package parasitics and bond wires respectively. The LC ladder approach appears to be adequate for the output network of a transistor, since the device parasitics are usually on the form of a LC ladder [3]. This is however not the case at the input of a transistor, where the parasitics are better modelled as an LC series resonance structure. In the latter case, a matching network of the band-pass type is better suited to absorb the device parasitics into the broadband impedance matching network (BIMN). Several works have already implemented such solutions [3], [4]. A simplified model of a transistor parasitics including the package [3] is shown in Figure 1.

In this work we present the design of a broadband RF power amplifier using an AFT27S006NT1 RF power LDMOS transistor from NXP. LDMOS is a reliable and well proven technology for RF transistors. These devices can deliver high output power and good efficiencies. Compared to GaN transistors, they possess high intrinsic gate and drain capacitances. This creates a challenging matching environment in which input and output impedances will be transformed to very low resistive impedance values for typical RF operation frequencies. This matching environment is similar to the scenario we face when designing amplifiers at frequencies of 30 GHz or above. Due to this reason, in order to explore the feasibility of broadband matching under this conditions, an LDMOS transistor was chosen for this design. A non-linear model for this this transistor was used for the design, although the model seemed properly represented only under small signal excitation conditions. Therefore, the whole matching network design was realized based on linear S-parameter simulations using this model. Despite what existing theoretical models like the one in Figure 1 suggested, we are making use of BIMN of the band-pass form for both, input and output of

the transistor, since the small signal optimum impedances obtained by examination of the S_{11} and S_{22} resemble in both cases a series LC structure for the device parasitics.

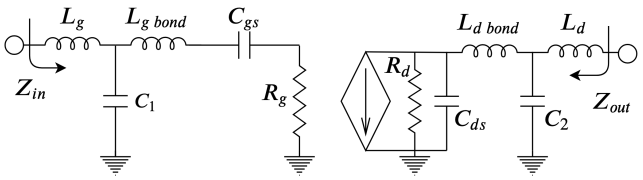


Fig. 1: Simplified model of a packaged transistor. C_{ds}, C_{gs} parasitic gate and source capacitances; $L_{d\,bond}$,$L_{g\,bond}$ bond inductances; C_1, C_2, L_g, L_d package ceramic capacitance and lead inductance respectively; R_g, R_d internal resistances at the input and output of the ideal device

II. Broadband Input Matching

A. Identify and model the input impedance

In order to design the input BIMN, the source impedance Z_S needs to be identified. The transistor was biased at drain and gate terminals and S_{11} was simulated right at the gate of the device. The simulated S_{11} for 1.4-2.6 GHz at the gate of the device is represented by the red trace of Figure 2. As can be appreciated, the impedance shown at the device input resembles that of a series RLC with a very small resistance $R = 0.0627 * 50 = 3.135 \approx 3.1\,\Omega$, which can be determined by denormalizing the value of the real impedance contour in which S_{11} is located. On the one hand, the resonance does not occur around the desired middle frequency of 2.1 GHz. On other hand, the central resonance frequency f_0 can be tuned by proper selection of the DC blocking capacitance to shift the resonance point at 2.1 GHz. In order to match the gate, a transformation from 3.1 to 50 Ω requires a large transformation ratio $n = 16.12$. Under this circumstances, a proper broadband matching at the device input was not realizable. Due to this issue and after testing several resistor values it was decided to implement a 10 Ω series resistor at the gate of the transistor to reduce the Q factor of Z_S, despite this will degrade the gain of the amplifier by several dB. Applying these modifications at the input of the device, the amplifier Z_S was shifted to the smith chart region represented by the blue trace in Figure 2.

In this situation, the input impedance was modelled by means of an resonance RLC with values of *R*=13 Ω, *C*=3 pF and *L*=2 nH. The S_{11} of this series RLC circuitry is shown in Figure 2 represented by the green trace. Due to the series resistance, the transistor was stabilized over the whole operating range, the stability for the low frequency region below 500 MHz was complemented by means of an additional low pass ladder of capacitors located in the gate bias lane with capacitor values of 6.8 pF, 27 pF, 240 pF and 1nF and 1 μF respectively. The stability of the amplifier was analyzed using the K-factor stability test [8], where the parameters k, μ_1, μ_2 and b_1 are calculated through the S-parameters of the amplifier. Unconditional stability of the amplifier will be achieved if $k > 1$ and $b_1 > 0$, whereas the condition involving the geometric stability factors $\mu_1, \mu_2 > 1$ is an additional indicator for amplifier stability. These parameters fulfilling

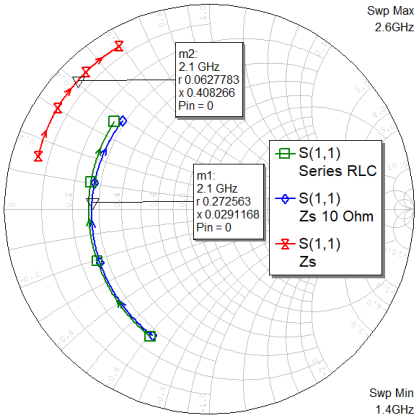


Fig. 2: Simulated S_{11} of the transistor input: Red-initial Z_S; Blue-Modified Z_S with a 10 Ω series resistor and a series capacitor at the gate of the transistor, Green-Equivalent RLC for Z_S

the unconditional stability condition are presented in Figure 3.

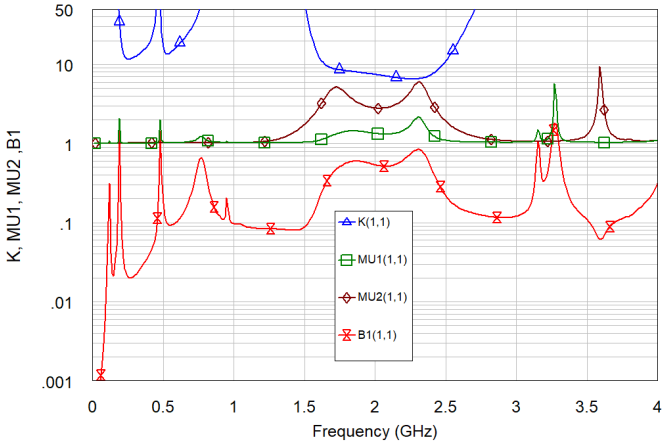


Fig. 3: Values of the K-factor stability test.

The RLC series resonance at the input of the device is then attempted to be absorbed by means of a band-pass BIMN. The matching network synthesis applied for the design of the input BIMN [4] is shown sequentially in Figure 4.

B. Synthesis of the input BIMN

The synthesis process makes the modelled RLC at the input of the transistor undergo a Band-pass filter (BPF) to low-pass filter (LPF) transformation, so that the series LC structure is converted to an equivalent inductor by means of the set of equations 1. Here L', and C' are the normalized values of L and C relative to R_{in} respectively, g_{req} is the normalized value of the equivalent LPF inductor, where ω_0 is the modelled RLC frequency of resonance and Δ is the aimed fractional bandwidth.

$$L' = L/R_{in} \tag{1a}$$
$$C' = C/R_{in} \tag{1b}$$
$$g_{req} = L'\omega_0\Delta \tag{1c}$$

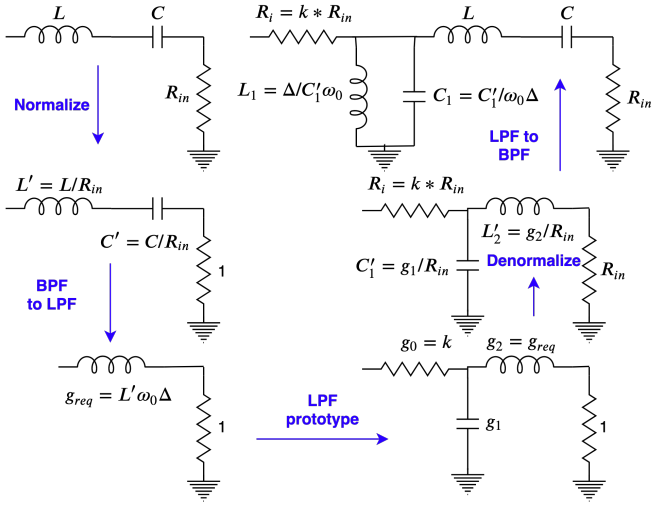


Fig. 4: Sequence for band-pass BIMN synthesis.

After calculating g_{req}, a LPF prototype with unequal termination is calculated making use of existing tabulated data [5], [6]. The order or the filter can be adjusted to fit the required g_{req} to be as close as possible to the value of the n^{th} element of the prototype g_n, which will then transform the initial input resistance R_{in} by a determined impedance transformation ratio k, in this work n=2 was selected. Once the LPF prototype is calculated, the process is reverted, first by de-normalizing the values by R_{in}, and then by making a LPF to BPF transformation as given by the set of Equations 2 and indicated in the last step of Figure 4.

$$C_1' = g_1/R_{in} \tag{2a}$$
$$L_2' = g_2/R_{in} \tag{2b}$$
$$L1 = \Delta/C_1'\omega_0 \tag{2c}$$
$$C1 = C_1'/\omega_0\Delta \tag{2d}$$

After this procedure, the device input parasitics are properly integrated into the band-pass BIMN, and the input resistance is transformed to a pure resistive value of $R_i = k * R_{in}$. The implementation of the shunt resonance LC of the band-pass matching network was accomplished by two symmetric half wavelength open stubs of characteristic impedance $Z_0 = 32\ \Omega$. At this point, an additional broadband impedance transformation from R_i to 50 Ω is required. For this purpose a broadband Chebyshev impedance transformer [7] was chosen in order to obtain the desired bandwidth by the real-to-real final impedance transformation. The calculated values of the BIMN synthesis for the values of the modelled RLC structure at the transistor input R = 13 Ω, L=2 nH and C=3 pF are summarized in Table I and the structure of the final input BIMN is shown in Figure 5.

TABLE I: Values for the input BIMN synthesis.

Δ	g_{req}	k	g_1	g_2	L_1(nH)	C_1(pF)	R_i
82	2.497	1.667	0.733	2.489	1.102	5.211	21.67

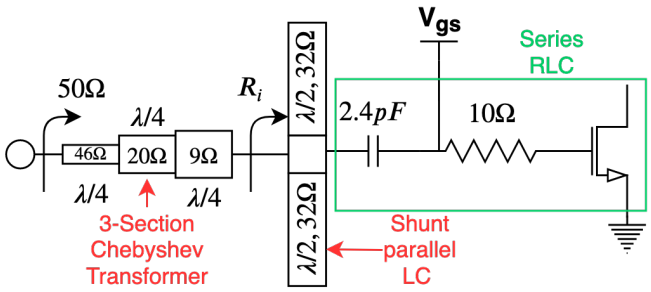


Fig. 5: Structure of the Input BIMN

III. Broadband Output Matching

A. Identify and model the output impedance

For the design of the output BIMN, the optimum load impedance $Z_{L\,opt}$ needs to be identified. This is usually achieved by simulating the load pull contours for power and efficiency under large signal excitation. However, the model available for this design only worked properly for small signal excitation. Therefore S_{22} is analyzed in order to identify the output impedance of the device at the drain terminal for the frequency range of interest. The simulated S_{22} resembles also in this case the response of a series RLC, although with a resonance frequency way above 2.1 GHz as indicated by the blue trace in Figure 6. To shift the impedance trajectory to the desired resonance frequency of 2.1 GHz, a series inductive transmission line was added at the output of the device, so that the resonance frequency is shifted to 2.1 GHz, the modified S_{22} is shown by the red trace in Figure 6. Finally, the output impedance is modelled again as a series RLC structure with R=3.9 Ω, C= 2.6pF and L = 2.2 nH. The frequency response of this series RLC is represented by the green trace in Figure 6, which models the output impedance of the device including the series inductive transmission line.

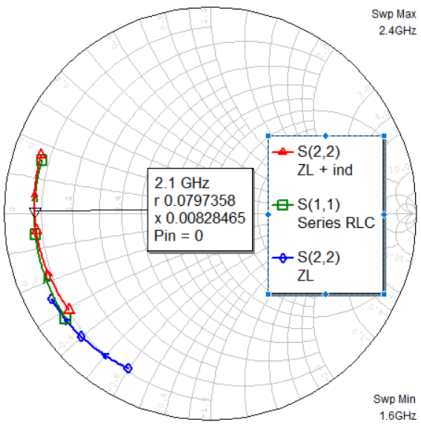


Fig. 6: Simulated S_{22} of the transistor output: Blue-initial Z_L; Red-Modified Z_L, Green-Equivalent RLC for Z_L.

B. Synthesis of the output BIMN

The synthesis method used for the design of the output BIMN is based on the work of Matthaei [1] and Dawson [2].

It consists on the synthesis of wide-band band-pass filters in the form of quarter-wavelength stubs and transmission lines between them, which produce the required resonance responses of the shunt parallel and series LC of a band-pass filter. The synthesis procedure for band-pass BIMN of order *n*=3 and *n*=4 have already been presented in previous works [3]. In this work we are implementing a *n*=4 band-pass BIMN, the correspondent design flow and equations are presented subsequently.

Step 1: Determination of the series RLC Q factor and calculation of the load decrement δ by means of equation 3 where Q can be derived either from L or C and R of the series resonance RLC, f_0 represents the central frequency of the band, and Δ represents the desired fractional bandwidth to be achieved.

$$\delta = \frac{1}{Q\Delta} = \frac{R}{2\pi f_0 L\Delta} = \frac{2\pi f_0 RC}{\Delta} \tag{3}$$

Step 2: Based on the work of Daswon [2], for *n*=4 the third-degree polynomial given by equation 4a needs to be solved to determine the parameters a and b, defined by equations 4b and 4c. The solution to this equation for *n*=4 is obtained by calculating c and r_4 as given by equations 4d and 4e. The values of $r_{4\,min}$, $r_{4\,max}$ and $r_{4\,mid}$ are shown in Table II as specified in the paper of Dawson [2].

$$16(xy)^3 + 8(xy)^2 + (2+8\delta)(xy) - (1+2\delta^2) = 0 \tag{4a}$$

$$x = sinh(a) = \sqrt{\left(\frac{c}{2}\right)^2 + r_4} + \frac{c}{2} \tag{4b}$$

$$y = sinh(b) = \sqrt{\left(\frac{c}{2}\right)^2 + r_4} - \frac{c}{2} \tag{4c}$$

$$c = 2\delta \sin\frac{\pi}{2n} \tag{4d}$$

$$r_4 = \frac{r_{4\,min} + r_{4\,max}\left(\frac{c}{r_{4\,mid}}\right)^2}{1 + \left(\frac{c}{r_{4\,mid}}\right)} \tag{4e}$$

TABLE II: Values of $r_{4\,min}$, $r_{4\,max}$ and $r_{4\,mid}$ for *n*=4 [2].

$r_{4\,min}$	$r_{4\,max}$	$r_{4\,mid}$
0.23031	0.25	1.008

Step 3: Calculate the normalized g_i values for each element in low pass Chebyshev form using the set of equations 5a to 5c. The necessary parameters D and $k_{j-1,j}$ are given by the set of equations 5d to 5f.

$$g_i = \frac{1}{\delta} \tag{5a}$$

$$g_j = \frac{1}{g_{j-1}(k_{j-1,j})^2}, j = 2,3,4 \tag{5b}$$

$$g_5 = \frac{1}{\delta D g_4} \tag{5c}$$

$$\theta = \frac{\pi}{2n} = \frac{\pi}{8}, n = 4 \tag{5d}$$

$$D = \frac{x}{\delta sin\theta} - 1 \tag{5e}$$

$$g(\theta) = sin\theta, f(\theta) = cos\theta \tag{5f}$$

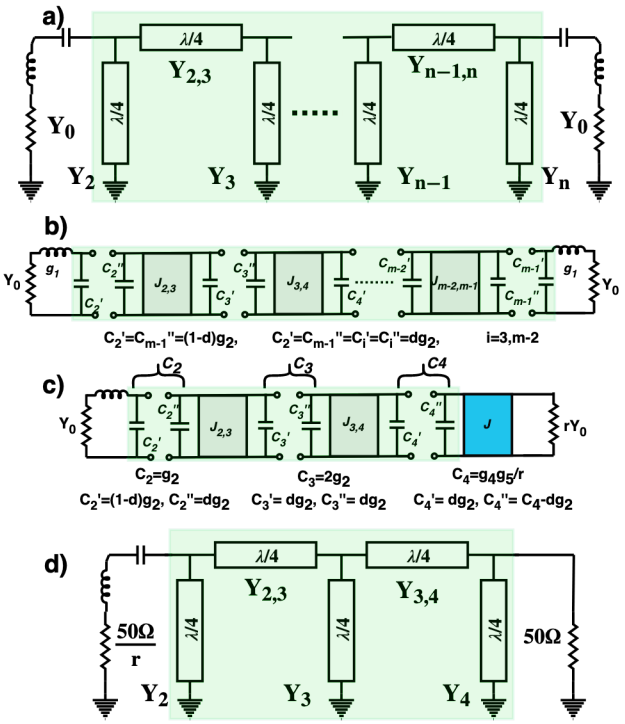


Fig. 7: Synthesis of the output band-pass BIMN. a) General band-pass filter structure with $\lambda/4$ resonators. b) Structure to derive synthesis equations for band-pass filter [1]. c) Structure to derive the synthesis equations for band-pass matching network. [3] d) Structure of the synthesized matching network using shunt $\lambda/4$ stub resonators.

$$k_{1,2} = \sqrt{\frac{g(\theta)^2 f(\theta)^2 + (f(\theta)^2 + D^2 g(\theta)^2) g(\theta)^2 \delta^2}{g(\theta) g(3\theta)}} \tag{5g}$$

$$k_{2,3} = \sqrt{\frac{g(2\theta)^2 f(2\theta)^2 + (f(2\theta)^2 + D^2 g(2\theta)^2) g(2\theta)^2 \delta^2}{g(3\theta) g(5\theta)}} \tag{5h}$$

$$k_{3,4} = \sqrt{\frac{g(3\theta)^2 f(3\theta)^2 + (f(3\theta)^2 + D^2 g(3\theta)^2) g(3\theta)^2 \delta^2}{g(5\theta) g(7\theta)}} \tag{5i}$$

Step 4: Calculate the impedance values of the $\lambda/4$ stubs required to implement the resonators of the band-pass filter structure depicted in Figure 7a as specified by the set of equations 6. These equations implement the synthesis method for a band-pass filter based on the work of Matthaei [1], whose procedure is summarized and illustrated in Figure 7b. The parameter d is a value ranging from 0 to 1 and might be freely chosen to adapt the impedance level at the center part of the filter by distributing the values of C_m in Figure 7b between C'_m and C''_m, w' is the normalized cutoff frequency of the original low pass prototype, and $C_\alpha = 2dg_2$. The set of equations 6 are valid up to *m*=7, which corresponds to a low pass prototype of order *n*=4.

$$Y_2 = Y_{m-1} = \frac{Y_0 w' C_2'}{g_0} tan\theta' + Y_0 \left(M_{2,3} - \frac{J_{2,3}}{Y_0} \right) \quad (6a)$$

$$Y_k = Y_0 \left(M_{k-1,k} - \frac{J_{k-1,k}}{Y_0} + M_{k,k+1} - \frac{J_{k,k+1}}{Y_0} \right) \Bigg|_{k=3,m-2} \quad (6b)$$

$$Y_{k,k+1} = Y_{m-k,m-k+1} = J_{k,k+1} \Big|_{k=2,m-1} \quad (6c)$$

$$\frac{J_{2,3}}{Y_0} = \frac{J_{m-2,m-1}}{Y_0} = \frac{1}{g_0} \sqrt{\frac{g_2 C_\alpha}{g_2 g_3}} \quad (6d)$$

$$\frac{J_{k,k+1}}{Y_0} = \frac{C_\alpha}{g_0 \sqrt{g_k g_{k+1}}} \Bigg|_{k=3,m-3} \quad (6e)$$

$$M_{k,k+1} = \sqrt{\left(\frac{J_{k,k+1}}{Y_0} \right)^2 + \left(\frac{w' C_\alpha tan\theta'}{2 g_0} \right)^2} \Bigg|_{k=2,m-1} \quad (6f)$$

$$\theta' = \frac{\pi}{2} \left(1 - \frac{\Delta}{2} \right) \quad (6g)$$

Step 5: The synthesized band-pass filter requires a modification in order to be converted to a band-pass BIMN terminated in a $Z_0 = 50\ \Omega$ impedance as shown in Figure 7c. Therefore, one half of the symmetric band-pass filter is removed and an additional admittance inverter J is implemented. This will modify the value of C_4 to the one required for a real impedance transformation, allowing the band-pass matching network to present a 50 Ω termination [3]. The value of C_4 is determined by assuming the admittance inverter to be $J = 1$, so that the capacitance C_4 provides the required susceptance for a real impedance transformation with the required impedance ratio r as shown in Equations 7a and 7b. The modified values of the admittance inverters and the admittance value of the $\lambda/4$ shunt stub Y_4 required for the BIMN synthesis after applying the aforementioned modification are given by the set of Equations 7. After acquiring the values of the admittance inverters given by Equations 7c and 7d. The modified final values of the $\lambda/4$ series and shunt stubs defining the final structure of the band-pass BIMN can be calculated as given by Equations 6a, 6b, 6c and 7e, this last being obtained by proper variable substitution [3] of C_4'' for C_2' and $M_{3,4}$, $J_{3,4}$ for $M_{2,3}$, $J_{2,3}$ on equation 6a.

$$\frac{J}{Y_0} = \frac{1}{g_0} \sqrt{\frac{r C_4}{g_4 g_5}} \quad (7a)$$

$$C_4 = \frac{g_4 g_5}{r} \quad (7b)$$

$$\frac{J_{2,3}}{Y_0} = \frac{1}{g_0} \sqrt{\frac{g_2 C_3}{g_2 g_3}} = \frac{1}{g_0} \sqrt{\frac{2 g_2}{g_3}} \quad (7c)$$

$$\frac{J_{3,4}}{Y_0} = \frac{1}{g_0} \sqrt{\frac{C_3 C_4}{g_3 g_4}} = \frac{1}{g_0} \sqrt{\frac{2 g_2 g_5}{r g_3}} \quad (7d)$$

$$Y_4 = \frac{Y_0 w' C_4'}{g_0} tan\theta' + Y_0 \left(M_{3,4} - \frac{J_{3,4}}{Y_0} \right) \quad (7e)$$

Step 6: As a final step on the output BIMN synthesis, the $\lambda/4$ shunt stubs were replaced by electrically equivalent $\lambda/2$ open stubs, whose admittance is determined by Equation 8. This conversion shown in Figure 8 was realized for the ease of lab tuning and fabrication and to avoid the non-idealities of a real short circuit in a printed circuit board, as well as to compare the performance by using both approaches of the BIMN. The final values of the BIMN are summarized in table III.

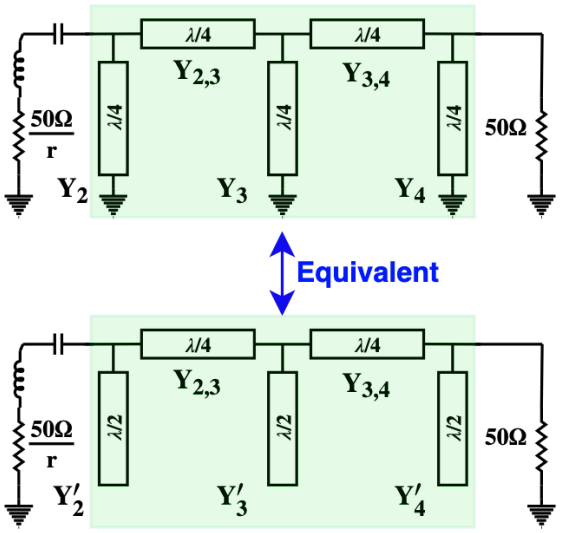


Fig. 8: Equivalent Structure of the synthesized matching network using shunt $\lambda/4$ or $\lambda/2$ open stub resonators.

TABLE III: Parameters for the BIMN synthesis.

f_0 (GHz)	2.1	$k_{3,4}$	0.9707	Y_2'	0.2517
δ	0.354	d	0.09	Y_3'	0.02697
Δ	0.381	θ'	1.272	Y_4'	0.01495
r	12.72	ω'	1	$M_{3,4}$	0.2248
g_0	1	$J_{2,3}$	0.0512	$Z_{2,3}$	19.5
g_1	2.824	$J_{3,4}$	0.02188	$Z_{3,4}$	45.7
g_2	0.7116	$M_{2,3}$	0.2892	Z_2	1.798
g_3	3.16	$Y_{2,3}$	0.0512	Z_3	17.35
g_4	0.3358	$Y_{3,4}$	0.02188	Z_4	30.26
g_5	2,323	Y_2	0.5563	Z_2'	3.973
$k_{1,2}$	0.7054	Y_3	0.05763	Z_3'	38.35
$k_{2,3}$	0.6668	Y_4	0.03305	Z_4'	66.88

$$Y_i' = \frac{tan^2\theta' - 1}{2 tan^2\theta'} Y_i \Big|_{i=2,4} \quad (8)$$

Some important considerations can be extracted by close analysis of the BIMN values in Table III. The value Z_2 for the $\lambda/4$ shunt stub connected directly to the device output presents a very small impedance close to 2 Ω, which would be practically unrealizable. For the electrically equivalent alternative of the $\lambda/2$ open stubs, the impedance Z_2' presents a value closer to 4 Ω, which might still be of unpractical dimensions. In order to increase the impedance of the open stub, a balanced version with two parallel stubs and a characteristic impedance close to 8 Ω was selected for the practical realization of the BIMN, which lead to still quite big, but at least realizable dimensions for the open stubs. Based on the aforementioned argumentation, the version with the $\lambda/2$ balanced open stubs, was the one selected for the final production of the PCB, whose output band-pass BIMN structure is highlighted in Figure 9.

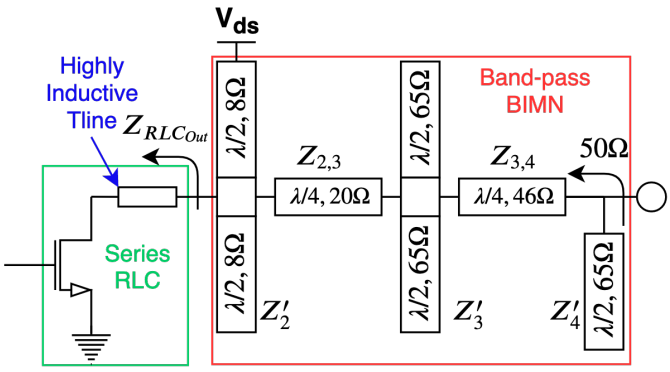


Fig. 9: Structure of the Output BIMN

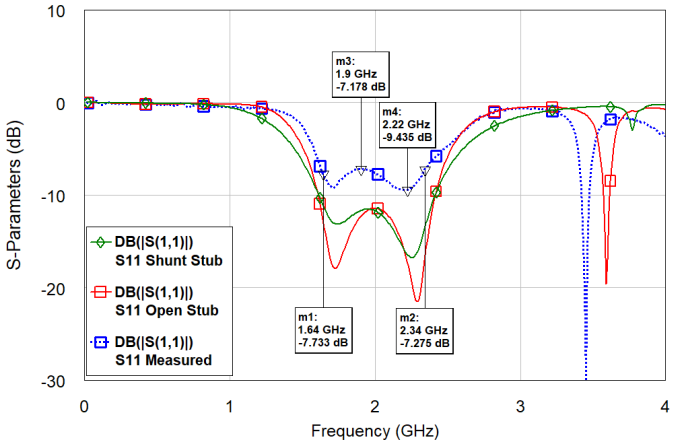


Fig. 10: S_{11} of the PA for the versions with $\lambda/4$ shorted stubs and $\lambda/2$ open stubs.

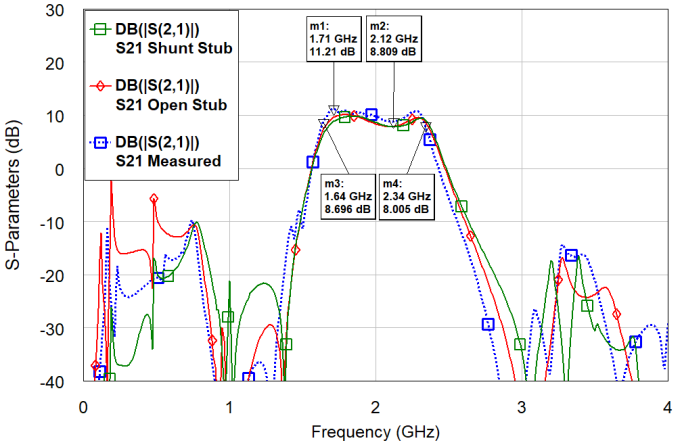


Fig. 11: S_{21} of the PA for the versions with $\lambda/4$ shorted stubs and $\lambda/2$ open stubs.

IV. Simulation and Measurements

Once the input and output BIMN in band-pass form are synthesized, the PA is fully EM simulated using AXIEM EM simulator from AWR simulation environment. A RO4350 substrate with 500 μm height and 17 μm conductor thickness was selected for the PA implementation on a PCB. The amplifier biasing circuit was designed by implementation of $\lambda/4$ lines at the central frequency of 2.1 GHz and a properly selected decoupling capacitor of 6.8 pF which will behave as a short circuit for the in-band frequencies. As a recap of what was previously mentioned in section II, further decoupling between DC and RF was implemented by a low pass ladder located right after the in-band decoupling capacitor with shunt capacitors of 27 pF and 240 pF, 1nF and 1 μF. In the gate bias lane, 10 Ω series resistors where also added to further improve stability for frequencies below 500 MHz without incurring into high losses. The amplifier is biased in class AB, with correspondent $V_g = 1.9\ V$ and $V_d = 28\ V$. The final structure of the amplifier was optimized and fine-tuned after EM simulation, therefore the final dimensions of each of the theoretically calculated components during the BIMN synthesis are adjusted and differ slightly from the initially predicted. The measurement setup consisted of two KEITHLEY 2400 source meters as well as a R&S®ZVL vector network analyzer. The performance of the designed PA was simulated and measured under small signal excitation, whose results are presented in Figures 10 and 11.

The simulated S_{11} of both amplifier structures using open and short circuited stubs are shown in Figure 10. Simulations show a value of S_{11} better than -10 dB for the frequency range of 1.6 to 2.4 GHz for both versions, with a total bandwidth of 800 MHz. Measured S_{11} of the fabricated version using open stubs is also represented by dotted lines in Figure 10. Measurements show a certain degree of degradation compared to the simulated S_{11}, where a maximum value of -9.435 dB is achieved as shown by marker *m4*. A value better than -7 dB is achieved for a 700 MHz bandwidth between 1.64 and 2.34 GHz, as highlighted by markers *m1* to *m3* in Figure 10. In an analogous manner, Figure 11 shows the simulated and measured values of S_{21}. Simulations show a maximum of 10.1 dB and a minimum of 7.9 dB small signal gain for the frequency range of 1.66 to 2.36, which corresponds to a gain ripple $\Delta_{Gain} = 10.1 - 7.9 = 2.2$ dB determined by the difference between the maximum and minimum values of gain inside the 700 MHz passband. The measured value of S_{21} presents a maximum of 11.21 dB and a minimum of 8.809 dB of small signal gain inside the passband as shown by markers *m1* and *m2* in Figure 11. Those values correspond to a gain ripple $\Delta_{Gain} = 11.18 - 8.806 = 2.401 \approx 2.4$, whereas a 700 MHz passband is determined by the frequency range of 1.64 to 2.34 GHz as shown by markers *m3* and *m4* in Figure 11. The measured value of S_{21} is in good agreement with the simulated S_{21}, showing a slight increase in the overall gain of 1 dB and the same predicted bandwidth of 700 MHz. Based on the previous considerations, a lower simulated bandwidth of 700 MHz was achieved for the output matching compared to the 800 MHz bandwidth achieved at the input matching. This result is not surprising, since a higher impedance transformation ratio was required at the output of the power amplifier in comparison to the input, limiting the amount of achievable bandwidth. In contrast to the simulated response, the measurements of the fabricated PA present the same 700 MHz bandwidth for both, output and input matching, while showing an improvement in S_{21} and a degradation of S_{11}. The aforementioned scenario seems to be

originated by a deviation of the gate resistor value from the $10\,\Omega$ used in the simulations. Lowering this resistor results in a rise of the amplifier gain while, at the same time, it reduces the input matching. The measured performance is still a great improvement in terms of bandwidth when compared to the performance shown in the transistor datasheet from NXP [9], where a very narrow bandwidth of around 100 MHz at 2.1 GHz is discussed, as can be seen in Figure 12. The counterpart of this increased bandwidth is the tradeoff in gain, since 12 dB of gain had to be sacrificed in order to obtain the wide bandwidth of the PA. Unfortunately, the non-linear model under which these simulations where developed did not work properly under large signal excitation, so further examination of the PA performance under large signal will have to be examined exclusively by measurements on the manufactured PCB of the power amplifier shown in Figure 13.

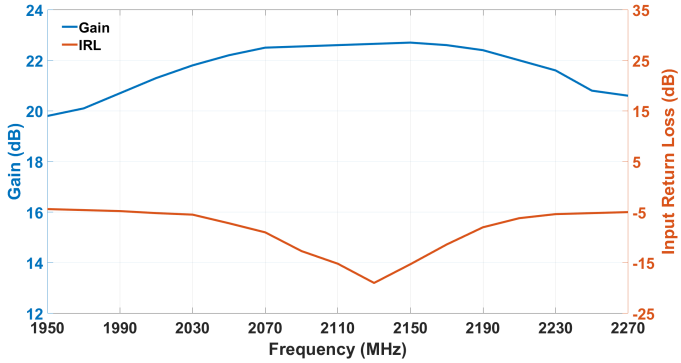


Fig. 12: S-parameters of the AFT27S006N RF power transistor at 2.1 GHz as depicted in the transistor datasheet from NXP [9].

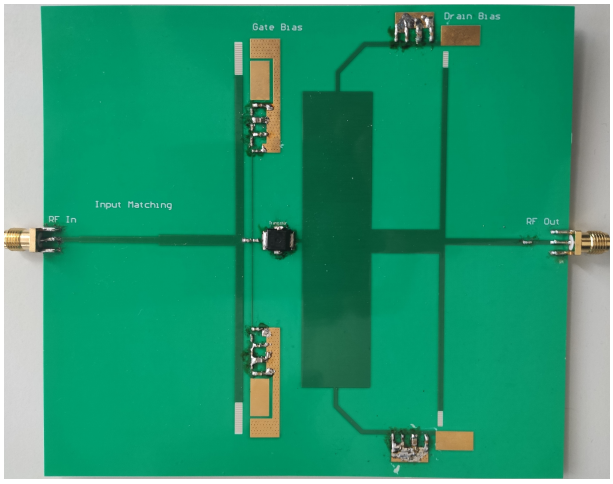


Fig. 13: Photo of the manufactured PCB.

V. Conclusion

In this work we have explored the feasibility of broadband matching a power amplifier based on an LDMOS transistor with very low real part of the input and output impedance if the parasitics at both sides are properly absorbed as part of the input and output matching networks. The parasitics at the input and output were modelled and properly adjusted to behave as an RLC series resonance circuit at the frequency of interest. There, reactive contribution was successfully absorbed by BIMN in band-pass form at the input and output of the transistor making use of two different approaches. The good agreement in terms of bandwidth between small signal measurement and simulation indicates that the size of the parasitic elements where properly identified during the design process. The price paid for this increased bandwidth of 700 MHz at 2.1 GHz center frequency is a degradation on the amplifier gain. This is caused by the series resistor at the gate of the transistor required to reduce the impedance transformation ratio to an acceptable value, which was an unavoidable condition for the BIMN synthesis method to be applied successfully. The degradation in the measured value of S_{11} could be compensated if more gain would be traded off by further reducing the impedance transformation ratio using a bigger series resistor at the gate of the transistor.

Acknowledgment

Special thanks to my colleages at IHF for their guidance and discussions in the field of power amplifiers during the initial period of my PHD. This work was my first amplifier design and by that time, it would have been a much harder task without their invaluable support.

References

[1] G. L. Matthaei, “Design of Wide-Band (and Narrow-Band) Band-Pass Microwave Filters on the Insertion Loss Basis” IEEE Transactions on Microwave Theory and Techniques, vol. 8, issue 6, pp. 580-593, November 1960.

[2] Dale E. Dawson, “Closed-Form Solutions for the Design of Optimum Matching Networks”, IEEE Transactions on Microwave Theory and Techniques, vol. 57, issue 1, pp. 121-129. January 2009.

[3] X. Meng, C.Yu and Y. Liu, “Design Approach for Implementation of Class J Broadband Power Amplifiers Using Synthesized Band-Pass and Low-Pass Matching Topology,” IEEE Transactions on Microwave Theory and Techniques, vol. 65 , issue 12, pp. 4984-4996. Dec. 2017.

[4] B. Gowrish, K. Rawat, A. Basu, S.K. Koul, “Broad-band matching network using band-pass filter with device parasitic absorption”, 82nd ARFTG Microwave Measurement Conference, November 2013.

[5] C. Bowick, J.Blyer, C.Ajluni, “RF Circuit Design,” 2nd edition, Newnes, 2007, pp.39-62.

[6] G. L. Matthaei, “Tables of Chebyshev impedance–transforming networks of low-pass filter form,” Proceedings of the IEEE, vol. 52 , issue 8 , pp. 939-963, August 1964.

[7] S.J. Orfanidis, “Electromagnetic waves and antennas,” Chapter 13, ECE Department, Rutgers University, November 2002.

[8] D.M. Pozar, “Microwave Engineering” 4th edition, Wiley, November 2011, Chapter 11.

[9] NXP Semiconductors Webpage. “https://www.nxp.com/products/rf/rf-power/rf-cellular-infrastructure/,”. Last visited March 26th 2020.

Manipulating Iron Filament with Permanent Magnets for FDM Printing for X-Band

Jan Köhler (B.Eng. M.Sc.)
University of Technology Graz
Institute of Microwave and Photonic Engineering (IHF)
Graz, Austria
jkoehler@tugraz.at

Wolfgang Bösch (Univ.-Prof. Dipl.-Ing. Dr.techn. MBA)
University of Technology Graz
Institute of Microwave and Photonic Engineering (IHF)
Graz, Austria
wbosch@tugraz.at

Erich Leitgeb (Ao.Univ.-Prof. Dipl.-Ing. Dr.techn)
University of Technology Graz
Institute of Microwave and Photonic Engineering (IHF)
Graz, Austria
erich.leitgeb@tugraz.at

Reinhard Teschl (Dipl.-Ing. Dr.techn.)
University of Technology Graz
Institute of Microwave and Photonic Engineering (IHF)
Graz, Austria
reinhard.teschl@tugraz.at

David Johannes Pommerenke (Univ.-Prof. Dipl.-Ing. Dr.-Ing.)
University of Technology Graz
Institute for Electronics (IFE)
Graz, Austria
david.pommerenke@tugraz.at

***Abstract*—This paper presents a novel technique for 3D printing that uses an external magnetic field to direct iron dust mixed into FDM filament. The mobility of the particles during the printing process is exploited to provoke an alignment by means of a static magnetic field. This method offers a completely new possibility to design passive high-frequency components. With very good agreement to the theory a method has been developed, manufactured and was measured within the X-Band.**

***Index Terms*—3D printing, Fused Filament Fabrication, iron filament, magnets, high frequency, passive components, X-Band**

I. INTRODUCTION

Revolutionizing the engineering world, FDM (Fused Deposition Modeling) or FFF (Fused Filament Fabrication) describes the manufacturing process of fusing materials layer by layer, whereby molten plastic is deposited into a grid shape by extrusion, heating, and application.
In a number of publications, the applicability of 3D printing for high-frequency technology is presented and completely new processes are demonstrated, like complex geometric structures of waveguides [1], planar filters [2], antennas [3], circuits. For example, the process of metallizing 3D-printed objects is often used, such as vacuum metallizing, electroplating or conductive painting [4]. In source [5] these methods were compared with each other and very good correspondence of the manufacturing processes with a possible applicability was found. Still all those discoveres are a compromise of used materials and tolerances in manufacturing.
In essence all those publications where devoted to creating popular devices from the technical literature, despite the capabilities of 3D-printing. Referring to CAD and to FEM simulations the aim is still pursued of having a strict separation of geometry and material, despite the capabilities of 3D-printing. By creating material embedded structures it is now possible to produce even more complex structure, like the mathematical concept of the fractal diamond [6] without the need of support structures.
The paper [7] goes into the specific application of other mechanisms in 3D printing, with a general discussion of the possible physical effects on materials and how to embed it. This paper though focuses on SLA printing and not on the more dominant FDM. Additionally, it is not addressing the possibility of arranging the magnetic field differently by their relative position of the poles. Being more interested on mechanical stress the paper does not address the possibilities of creating a novel type of passive high-frequency components, like embedding it.

In this paper a new FDM process is presented by using iron filament along with a static magnetic field. Polylactide (PLA) enriched with iron dust provides the basis for this concept. Typical for the FDM process, the iron filament is being heated to be positioned within the printing process, making in parallel the iron particles also mobile. Using a static magnetic field within the printing procedure, the iron is being re-arranged along the given field lines. Cooling the filament immediately after the extrusion, the iron particles are being fixed into a their position, creating an embedded structure. This novel method is being investigated and the capabilities of its use for high frequency components.

For this paper, the manufacturing process was developed, built, refined, with measuring its samples. The evaluation of the measurement results corresponds very well with this theory.

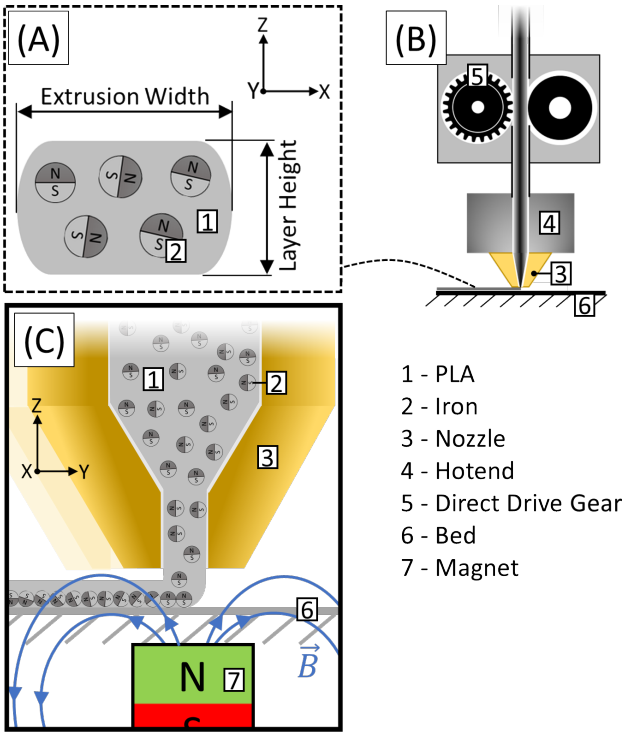


Fig. 1. FDM printing with permanent magnets as (A) cross section of filament with iron, (B) structure of extruder for 3D-printer, (C) close-up of nozzle

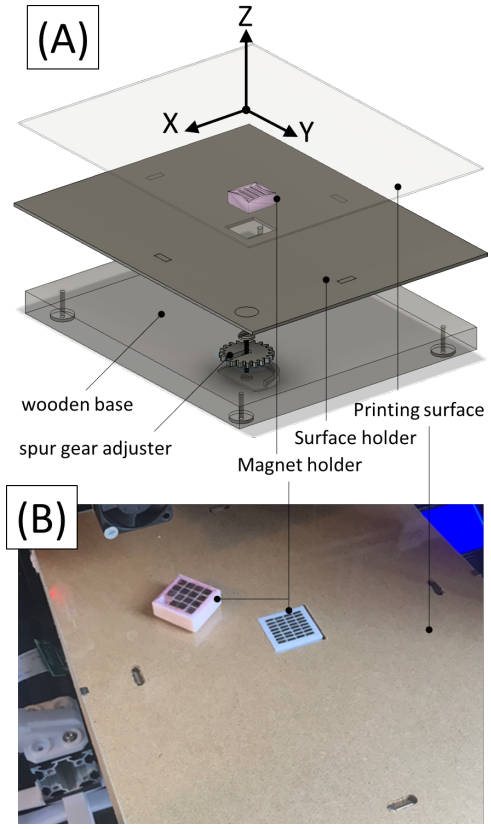


Fig. 2. Designed equipment for printing process with magnets (A) design and components, (B) manufactured design

II. DESIGN OF A SAMPLE ITEM

A. Describing the structure

The FDM printer used is a Cartesian construction type, whereby the respective axes X, Y, Z can move independently from one other in only one direction each. The fourth axis is the so-called extruder, being only responsible for feeding the filament and is the collective term for all components involved (i.e. nozzle, hotend, gearing). A direct drive extruder was used, whereby the distance between nozzle head and motor is kept small in order to achieve more controllability of the printing process and thus having higher accuracy for manufacturing. In the process shown in Fig. 1 filament is transported into the hotend by means of the integrated mechanism, melted through heat[1], pressed through the nozzle and positioned on the bed. Vertical structures are created by repetitive layering of the described process. In parallel for this particular filament, the containing iron particles also become physically mobile, as shown in Fig. 1 (A), (C). Each individual iron particle has the properties of a dipole and thus aligns itself to external magnetic field. This alignment is deliberately provoked by inserting permanent magnets into the printing bed. A spatial separation is given by an additional layer (e.g. glass, acrylic glass, glass fiber fabric).

As shown in Fig. 2 (A), great care was taken to use non-magnetic materials (e.g. wood, glass, paper, PLA).The base was build from a plywood panel on which a cardboard mesh plate is mounted, which in turn holds a glass plate or any other carrier. A gap was created within cardboard mesh plate, to position the magnetic holders (Fig. 2 (B)). Vertical adjustment were ensured with four spur gear adjuster, integrated at the corners. To ensure not having any vertical offset, an integrated mechanism is used to scan the surface [2], creating a 5×5 matrix and automatically adding it to the print job, creating a constant reproducibility of the samples.

For visualization purposes of the proposed process, the silhouette of the entry of an X-Band waveguide will be used (Fig. 3 (A)). The dimensions of the contours of the waveguide are 22.80 and $10.14\,mm$ [3]. To recreate the silhouette with static magnetic fields, a holder for the neodymium permanent magnets was printed from PLA, in which four bar magnets ($8 \times 4 \times 3\,mm$) and four cuboid magnets ($10 \times 10 \times 2\,mm$) could be fixed (see Fig. 3 (B)). The arrangement of the permanent magnets was chosen so that the same poles pointed to the center to form a cube (Fig. 3 C)). Making the magnetic field visible, a special foil was placed on the carrier with magnets in place (Fig. 3 (D)), which is used repeatedly in this paper. Through the repulsion of the magnetic field to the center of the holder the silhouette of the waveguide is being formed, with the iron particles preferably accumulating along the center of the magnet along the center of the poles. In order to check whether the same illustration could also be reproduced using the 3D printing, iron filament was used to printed on top of magnet carrier, as illustrated in Figure 3 (E). By creating a very thin layer making the sample translucent and shining light onto the sample, the corresponding magnetic field is illustrated. This effect should be further investigated with

[1]Depending on the material, approx. 200°C

[2]BLTouch V1.1

[3]Manually measured values with caliper gauge

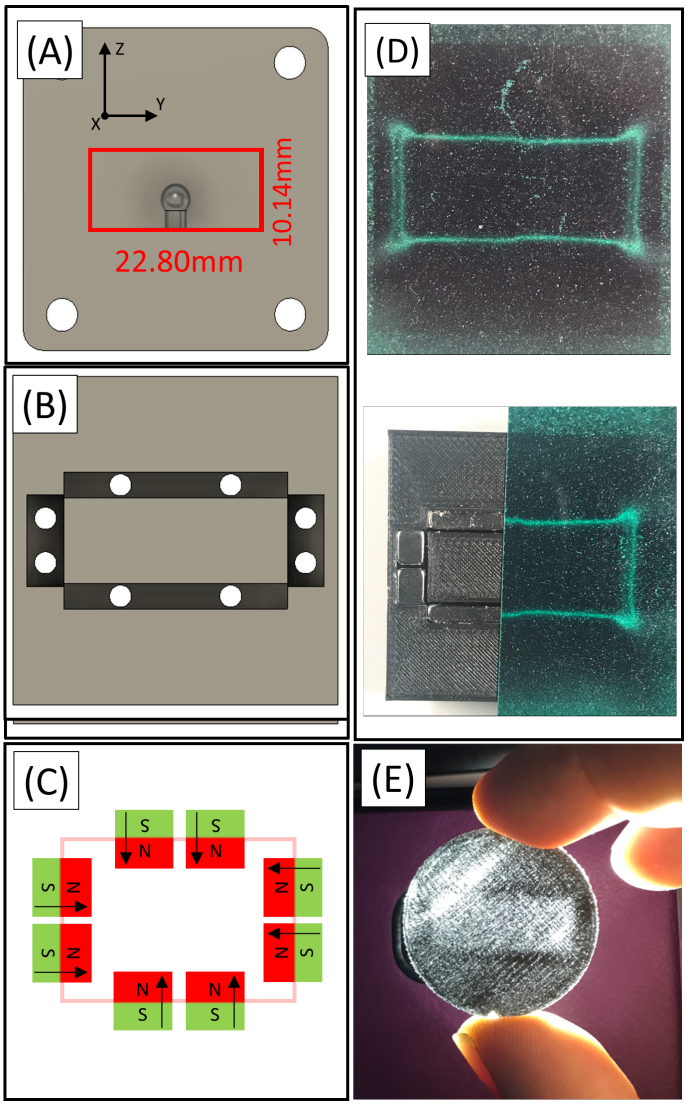


Fig. 3. Production process with iron filament and magnets (A) dimensions of waveguide, (B) Top view of magnetic holder, (C) Alignment of the magnets used, (D) Field distribution using magnetic foil, (E) printed sample

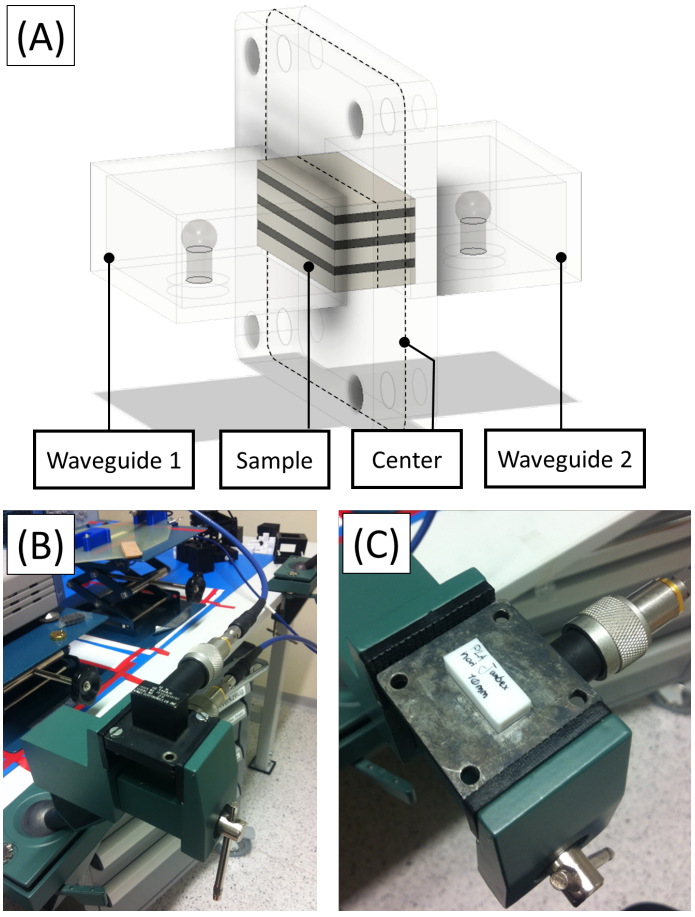


Fig. 4. Measurement setup for samples (A) Waveguide and with sample positioned, (B) composite waveguides , (C) single waveguide with installed sample

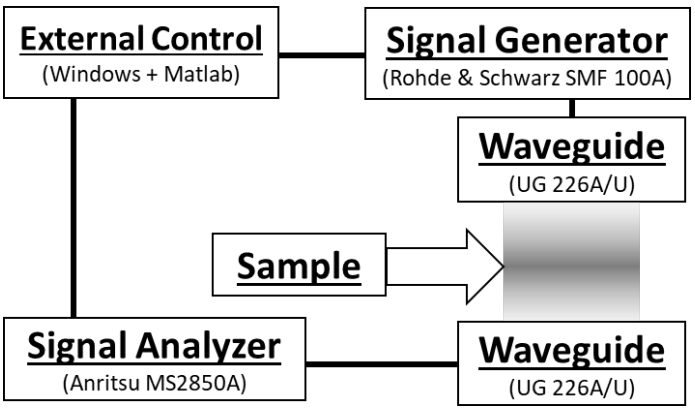


Fig. 5. Measurement setup used

measurements to determine the influence.

B. Production and Measurement

The proposed manufacturing process shall now be investigated, determining the influence on radio frequencies. To have non obstructing boundary conditions the samples were created so it could be placed flush in two connected waveguides (Fig. 4 (A)).

Setup: The measurement setup consists of a signal generator, a connected signal analyzer and an external control (Fig. 5). The inserted samples were clamped into the middle of the waveguide (Fig. 4 C)) calibrating it with calipers [4]. This structure was connected to a second identical waveguide.

Preparation of the samples: Two filament types were used, white PLA and iron filament from Protopasta. To ensure consistency and reproducibility each filaments was calibrated in prior. The printer was also calibrated independently (e.g. stepper motor setting, proportional-integral-derivative control (PID), horizontal size compensation). A nozzle diameter of $0.4\,mm$, a layer width of $0.48\,mm$ and a layer height of $0.2\,mm$ was used. All samples were prepared with an outer shell of two layers and the core was filled with a density of 95% [5] with a diagonal pattern. All samples were geometrically measured after manufacturing and a deviation of $\pm 50\,um$ was determined.
The influence on the high-frequency signal is to be investigated on the basis of the material properties of the filament mixed

[4] accuracy $10\,um$

[5] This is equivalent to a solid infill

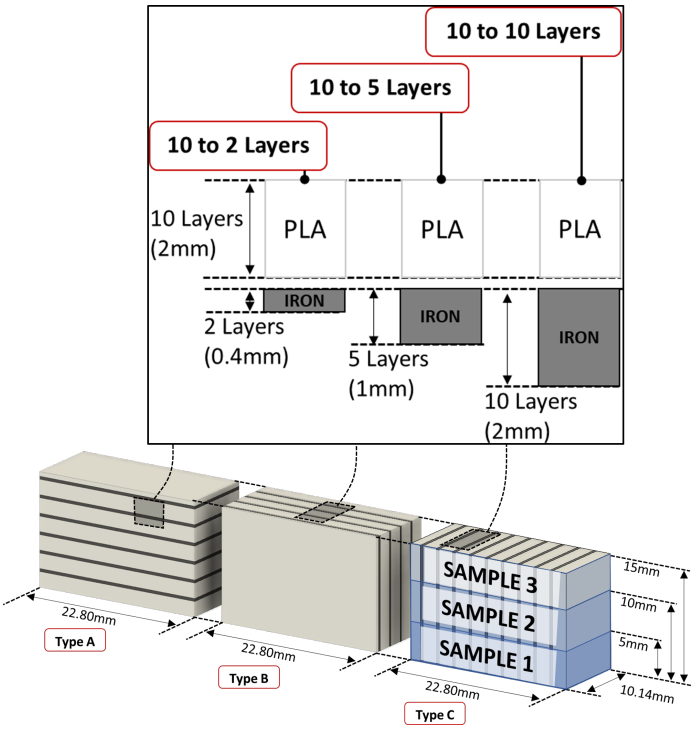


Fig. 6. Change parameters based on layer ratio and height

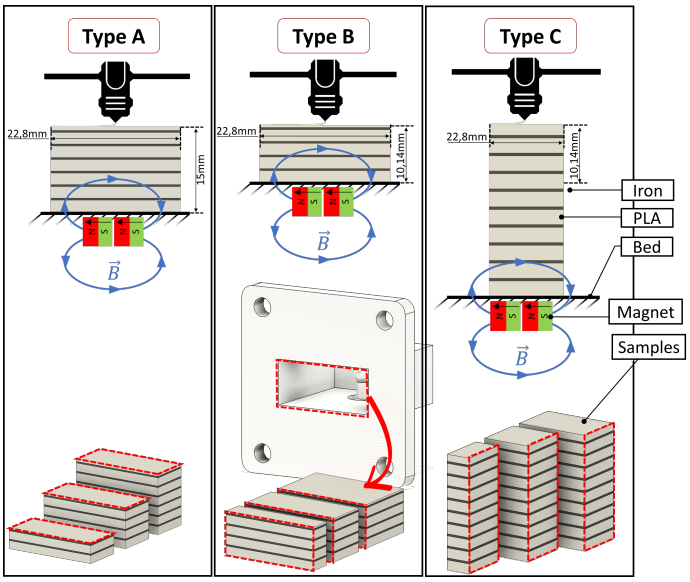


Fig. 7. Printing method for samples for Type A, Type B and Type C with magnet in bed

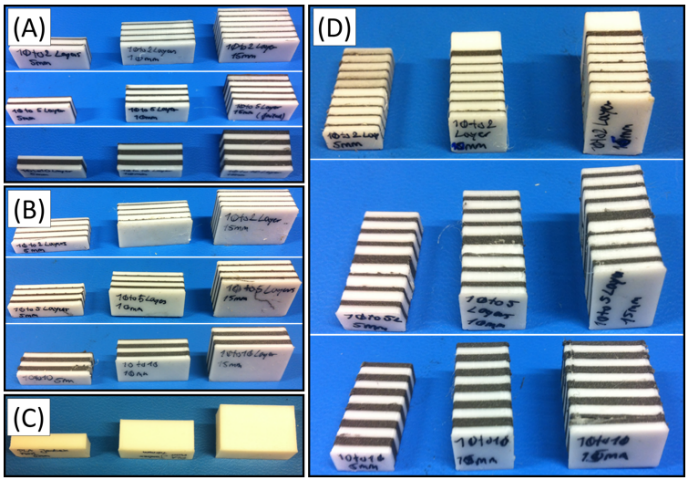


Fig. 8. Printed samples for (A) Type A, (B) Type B, (C) bare PLA, (D) Type C

with iron dust. For this purpose, layered samples were prepared, which consist partly of PLA and iron filament (Fig. 6). The density is controlled by the layering method of the 3D printer and the proportions are controlled by the layer height. According to source [8] the PLA has a permittivity of $\varepsilon_r = 2.7$ and should be kept constant in the samples. According to the manufacturer, the iron filament has a permeability of $\mu_r = 7-10$. As illustrated in Figure 6, the PLA was maintained at a constant height of $2\,mm$ (10 layers), with the height of the iron filament being varied between $0.4\,mm$ (2 layers), $1.0\,mm$ (5 layers) and $2.0\,mm$ (10 layers). These samples were made in three variations, $5\,mm$ (sample 1), $10\,mm$ (sample 2), $15\,mm$ (sample 3). Furthermore, three directions of the layers were created, relative to the used waveguide. The alignment were chosen as follows; for Type A orthogonal (Fig. 8 (A)), for Type B horizontal parallel (Fig. 8 (B)) and for Type C vertical parallel (Fig. 8 (D)). In order to be able to observe the influence of the iron, samples were also made out of only PLA for comparison (Fig. 8 (C)). A selection of the samples produced is shown in Figure 8 (A)-(D).

Preparation of the samples with magnets: Following the same procedure and as described in the section II-B, further samples were prepared using the method described, embedding neodymium magnets into the bed, as illustrated in Fig. 7. The aim of this is to imprint structures into the object with help of iron particles, by changing its density. Two types of magnets were used; $5 \times 5 \times 5\,mm$ (Magnet 1, $11.8\,N$) and $12 \times 12 \times 12\,mm$ (Magnet 2, $61.8\,N$). For these specific magnets holders were made out of PLA for positioning (Fig. 9 (Top)). Two holders have been designed for magnet 1. As illustrated in Figure 9 (B)(left), a chessboard-like distribution was chosen, with the polarization being distributed vertically inverted. The resulting field distribution of this forms a grid (Fig. 9 (C)(left). As can be seen in Fig. 9 (B)(middle), the same magnets were secondly formed into a ring structure with the same polarization in the center and opposite polarization at the edges. The resulting field distribution forms a square ring in the center. Correspondingly, magnet 2 was formed into a 2×2 matrix with rectified polarization for the third holder. It should be emphasized that magnet 2 has 5.2 times more force compared to the first magnet.

Measurements: The measurements are representing the signal transmission between two waveguides, with an inserted sample as an interference. In the first series of measurements, Type A samples with different layer height ratios were compared with each other, *10 to 2*, *10 to 5* and *10 to 10 layers*. A full uninfluenced transmission corresponds to 100% in these considerations. To establish causality to the influence of the iron, all measurements are compared with those from pure PLA.
The height of the samples was changed from $5\,mm$, to $10\,mm$, to $15\,mm$. Figure 10 (A) shows sample Type A, with a height of 5mm. The transmission of the signal experiences an attenuation

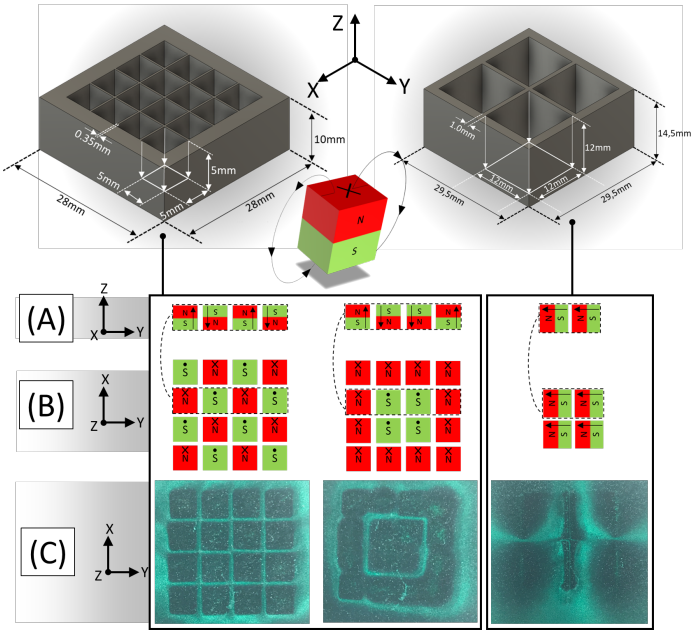


Fig. 9. Magnetic holders used for experiment with installed magnets as (A) side view, (B) top view, (C) top view with magnetic foil

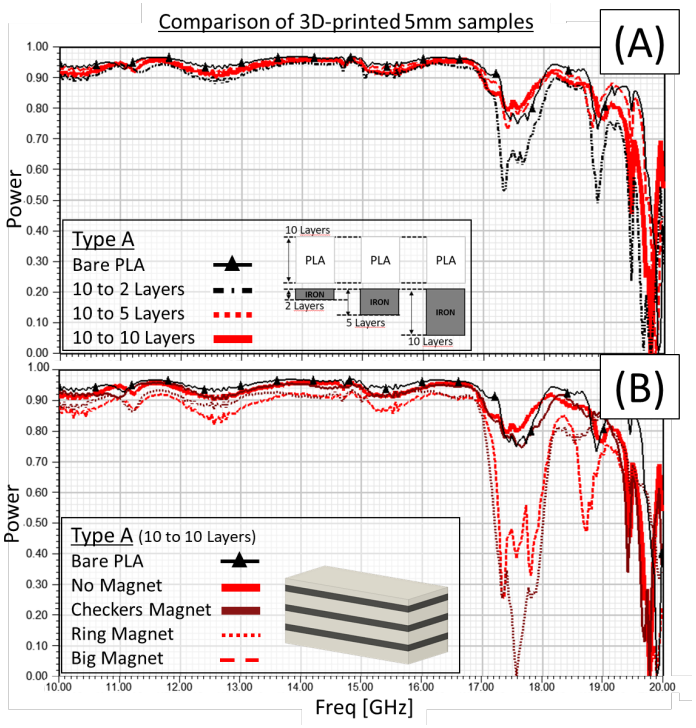


Fig. 10. Measurement for sample Type A $5\,mm$ with (A) change of layer ratio, (B) magnet used in production process

with a transmission value of 90-95% from 10 to $17.5\,GHz$. Noticeable in the sample for *10 to 2 layers* are the frequency points $17.5\,GHz$, $18.8\,GHz$, with a value of 60-65%. This sample has the most material transitions between PLA and iron (see Fig. 8 (A). In Fig. 11 (A) shown for sample Type A, with a height of $10\,mm$, a stronger damping and fluctuation can now be observed with a transmission of 80-95%. Striking frequency points here are 11, 14, 16.4 and $17.8\,GHz$. The sample *10 to 2 layer* experiences the strongest attenuation in this measurement. Striking points for these are $16.4\,GHz$ at 70% and $17.3-18.4\,GHz$ at 46-59%.

In Fig. 12 (A) shown for sample Type A, with a height of $15\,mm$, a further increase in damping can be observed. From $10-16.5\,GHz$ the value fluctuates from 70-95%. Striking points here are 11 and $15\,GHz$. At $17.9\,GHz$ signal transmission is no longer possible and at $19.4-20\,GHz$ the value varies from 20-75%. It can be assumed that the number of transitions between the two materials used in the samples affects the attenuation of the signal transmission.

Now a batch of identical samples is being compared, but with integrated magnets within the printing process. In order to be able to draw conclusions about the value, *10 to 10 layers* are used, since most iron filament is used. The graphs of Fig. 10 (B) of the measurements show fluctuations in the curves. As it can be seen from the graph Fig. 10 (B), the value of $10-17\,GHz$ varies from 82-96%. Striking points are the ranges $17.0-18.0\,GHz$ and $18.7\,GHz$. For the sample Type A *10 to 10 layer Big Magnet* the value varies from 25-55%. For sample Type A *10 to 10 layer Ring Magnet*, signal transmission is not possible at $17.6\,GHz$. The same applies to all samples at $> 19.5\,GHz$.

For further consideration sample Type A *10 to 10 layers* for 10mm in Fig. 11 (B) is being observed. Comparing the results to the samples without the magnets, more fluctuation is being observed. The samples for the *Checkers Magnet* varies at $10-16.4\,GHz$ with 74-88%, while for the *Ring Magnet* an improvement of the transmission of 92-96% can be observed. Striking points here are 11, 14, 16.6, $17.8\,GHz$, whereby the values are 40-75%. For the sample with *Checkers Magnet* a missing transmission at $17.8\,GHz$ can be detected.

Finally, this should be compared with Fig. 12 (B) Sample Type A, *10 to 10 layers*, with a height of $15\,mm$. From $10-17\,GHz$ there is less fluctuation compared to the previous two values. There are conspicuous places 10.5, 15.0 and $17.5\,GHz$, where a change of 5% can be observed, partly in additional attenuation but also improvement. It is possible that the size of the test sample has an influence on the signal transmission in the waveguide. Furthermore, at $19\,GHz$ an improvement of the transmission of 8% can be seen for *Ring Magnet* and *Big Magnet*.

Measurement with magnets: Considering these first results, it can be assumed that the iron particles are influencing the signal transmission. To support this thesis, the sample Type B and C are being compared. Here again the layer ratio *10 to 10* is chosen, with a height of $5\,mm$ for sample Type B (Fig. 13 (A)) and sample Type C (Fig. 14 (A)). Both show low fluctuations in the range of 10 to $17\,GHz$, regardless of the magnets used. The striking point here is $17.3\,GHz$ for both samples. An improvement in transmission can be observed for both Type B and C. Compared to the magnet-free counterparts in Type A, an improvement of 35% in signal transmission for *Checkers Magnet* and *Big Magnet* can be detected. Type B *Checkers* also shows 35% improvement. A further improvement can also be seen in 18.1 and $19\,GHz$. This property can also be found in the next higher iteration, for Type C $10\,mm$ Fig. 14 (B). Another improvement of the transmission within 10-15% for all magnet variants can be found at $10-17\,GHz$. *Ring*

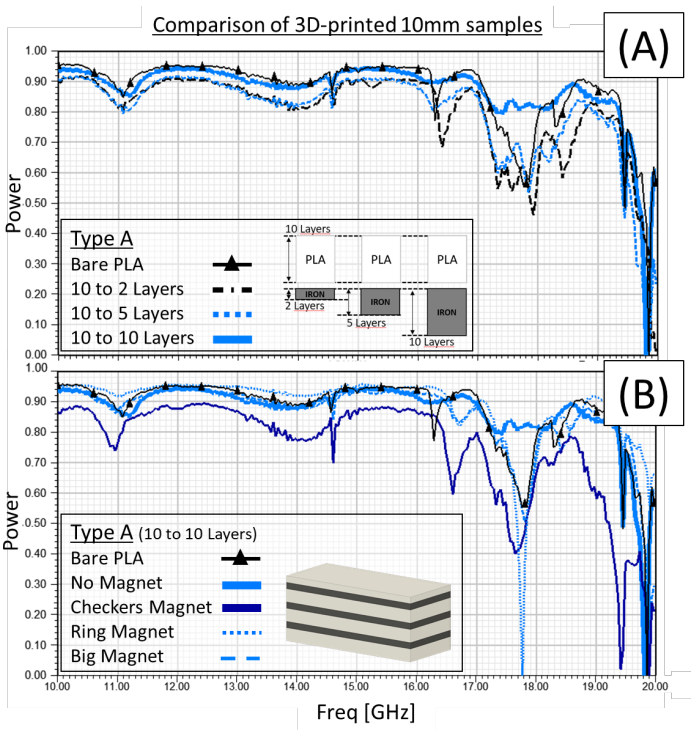


Fig. 11. Measurement for sample Type A $10\,mm$ with (A) change of layer ratio, (B) magnet used in production process

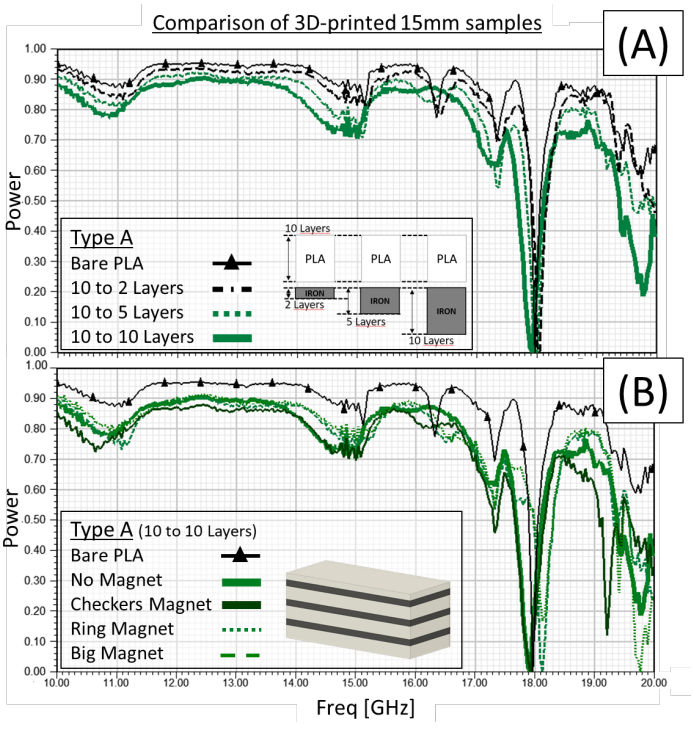


Fig. 12. Measurement for sample Type A $15\,mm$ with (A) change of layer ratio, (B) magnet used in production process

Magnet and *Big Magnet* are showing an improved transmission from 10 to $17\,GHz$. The *Checkers Magnet* always follows the curve of the magnet-free type. Again at $17.5\,GHz$ all types have little transmission. For *Big Magnet* the value is reduced to 42% and for *Checkers Magnet* the transmission is cutting off at $17.6\,GHz$. Similar characteristics can also be found in the graph of Fig. 13 (B), with Type B $10\,mm$. An improvement in signal transmission of 10-15% can also be observed in the $10 - 16.5\,GHz$ range, but only for *Checkers Magnet* and *Big Magnet*. *Ring Magnet* shows a deterioration of the transmission from 10-30%. The lowest signal transmission applies to all types at $17.5\,GHz$. The low signal transmission of the magnet-free type is additionally amplified by the magnetic antagonists.
Comparing the last iteration of Type B $15\,mm$ of Fig. 13 (C). A slight change in the signal transmission can be observed with the magnetic imprints. Only the *Big Magnet* version shows 10-20% improvement. In addition, there is no complete attenuation, as it is the case for all the other samples. This similar property can also be seen in Type C Fig 14 (C). All curves are similar to the magnet-free type, with only the *Checkers Magnet* variant showing 10-15% improvement at $11.8\,GHz$ and $17.5\,GHz$. Since this is a recurring phenomenon for this type, regardless of the composition, it can be assumed that the height of $15\,mm$ influences the results. This should be further investigated by using longer waveguides and creating additional iterations of heights for the samples.

III. CONCLUSION

The aim of this paper was to investigate whether the electromagnetic properties of filaments with iron dust can be manipulated when they printed under the influence of static magnetic fields within the FDM printing process. We illustrated the possibility of using the 3D printing process and modifying it in this way, by which embedded neodymium magnets change the position of iron dust within the used filament. Samples were created with this proposed purpose and compared with their magnetic free counterpart. The proportions were controlled with the typical FDM layering mechanism. The measurements were performed over a frequency range of $10 - 20\,GHz$ by inserting the samples between two interconnected waveguides. The measurements provided information on the degree of attenuation of the transmitted signal power, as well as how and whether the manipulation with static magnetic fields generates influence. The measurements confirmed this theory. Based on examples, even an improvement of the signal transmission up to 20% could be generated compared to the untreated, magnetic free counterpart. At the same time, this method also caused additional attenuation of the signal in other examples. From this it can be concluded that the wide range of modelling possibilities results in an extensive range of applications and thus represents a novelty.

ACKNOWLEDGMENT

The authors acknowledge the collaboration from Prof. David Johannes Pommerenke, Dipl.-Ing. Dr.techn. Reinhard Teschl and Dipl.-Ing. Kai Parthy of Lay-Filaments their help in the fabrication process and measurements and for valuable discussions about the experiments.

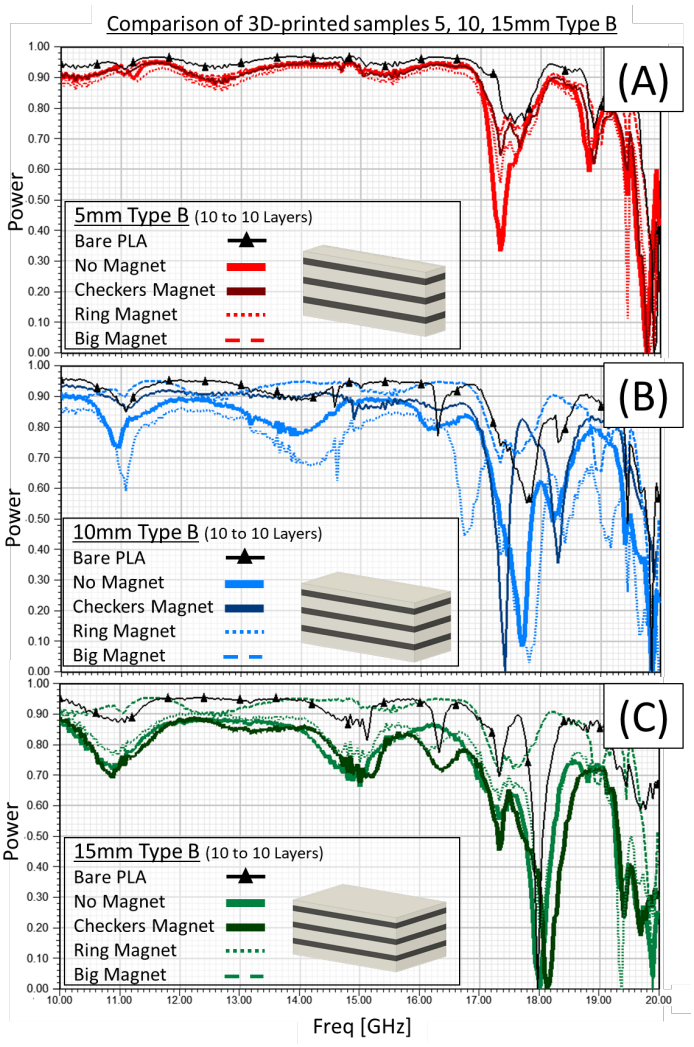


Fig. 13. Measurement for sample Type B *10 to 10 layers* with magnets (A) height $5\,mm$, (B) height $10\,mm$, (C) height $15\,mm$

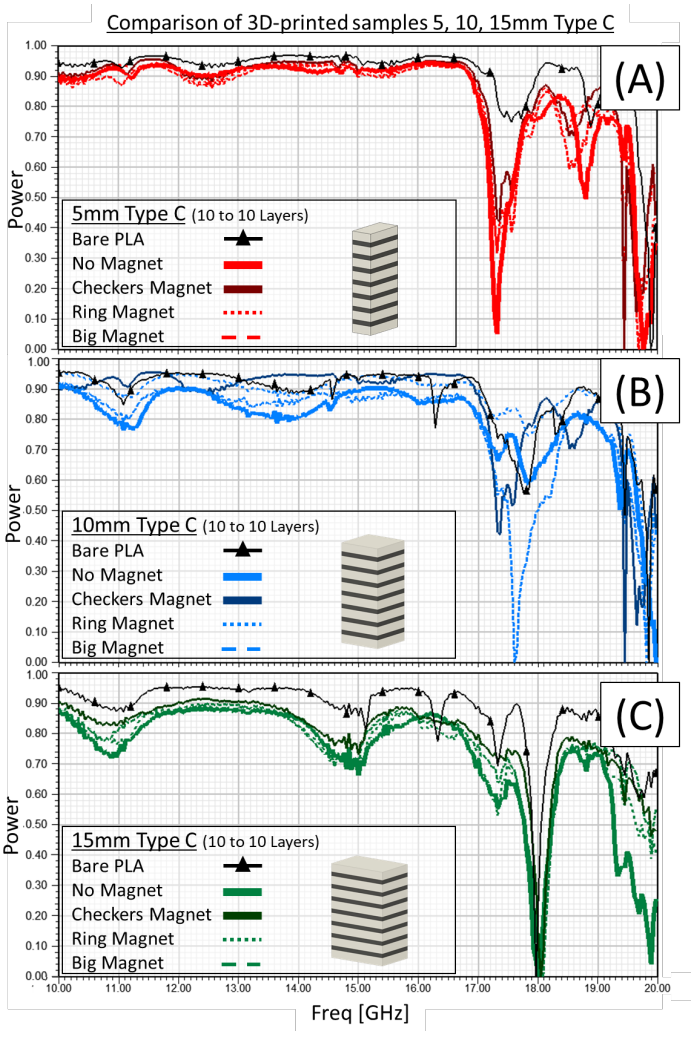


Fig. 14. Measurement for sample Type C *10 to 10 layers* with magnets (A) height $5\,mm$, (B) height $10\,mm$, (C) height $15\,mm$

REFERENCES

[1] *2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting*. IEEE, 2019.

[2] *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*. IEEE, 2018.

[3] *2015 European Microwave Conference (EuMC)*. IEEE, 2015.

[4] G. Limodio, Y. de Groot, G. van Kuler, L. Mazzarella, Y. Zhao, P. Procel, G. Yang, O. Isabella, and M. Zeman, "Copper-plating metallization with alternative seed layers for c-si solar cells embedding carrier-selective passivating contacts," *IEEE Journal of Photovoltaics*, vol. 10, no. 2, pp. 372–382, 2020.

[5] K. V. Hoel, S. Kristoffersen, J. Moen, K. G. Kjelgard, and T. S. Lande, "Broadband antenna design using different 3d printing technologies and metallization processes," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2016, pp. 1–5.

[6] *2017 IEEE MTT-S International Microwave Symposium (IMS)*. IEEE, 2017.

[7] M. Roy, P. Tran, T. Dickens, and A. Schrand, "Composite reinforcement architectures: A review of field-assisted additive manufacturing for polymers," *Journal of Composites Science*, vol. 4, no. 1, p. 1, 2020.

[8] *2016 22nd International Conference on Applied Electromagnetics and Communications (ICECOM)*. IEEE, 2016.

Comparison of Radiation Exposure between DVB-T2, WLAN, 5G and other Sources with Respect to Law and Regulation Issues

Peter Mandl
Institute of Microwave and Photonic Engineering
Graz University of Technology
Graz, Austria
peter@mandl.org

Pirmin Pezzei
Institute of Microwave and Photonic Engineering
Graz University of Technology
Graz, Austria
pezzei@tugraz.at

Erich Leitgeb
Institute of Microwave and Photonic Engineering
Graz University of Technology
Graz, Austria
erich.leitgeb@tugraz.at

***Abstract*— Current trends show that the demand for Internet bandwidth, especially regarding mobile end devices, is increasing significantly and will continue with an exponential growth in the future. In this regard, it is necessary to provide new technologies to meet the bandwidth requirements. In particular, the technology standard 5G and others are being promoted and realised worldwide, which rises a public discussion regarding the relevant non-ionizing electromagnetic radiation exposure for the population. In view of this field of tension, this publication presents a comparison of the non-ionizing electromagnetic radiation exposure between mobile radio transmitters, state of the art TV broadcast transmitters like DVB-T2 and WLAN base stations as part of long-term measurements. In particular, the different power flux densities of the technologies mentioned are measured, compared and also discussed with regard to the legal framework and limits.**

Keywords—Mobile phone base station radiation measurements (2G, 3G, 4G); 5G Next Generation Networks; DVB-T2; WiFi; WLAN; legal electromagnetic radiation limits; health electromagnetic radiation limits; non-ionizing electromagnetic radiation

I. INTRODUCTION

It is a proven fact that the bandwidth requirement increases exponentially and will continue to grow due to the extremely rising number of mobile Internet users, see Fig. 1.

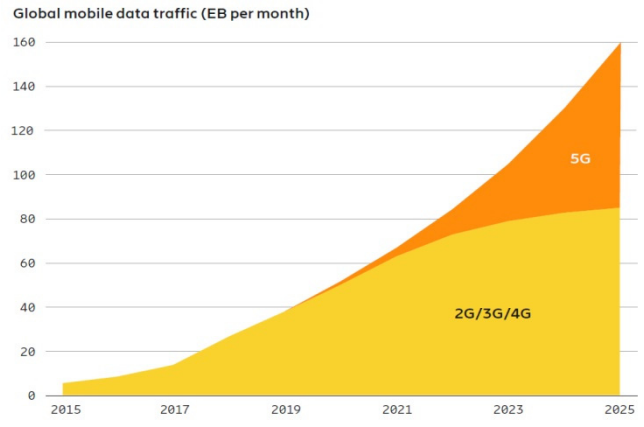


Fig. 1 Global mobile data traffic (EB per month) [1]

In the near past, different media have been used to identify possible health hazards from cell phone technologies, especially from the new 5G technology. To cite just one curious example, arson attacks were carried out on 5G masts in April 2020, especially since it was suspected that these mobile masts could be linked to COVID-19, see [2] and [3].

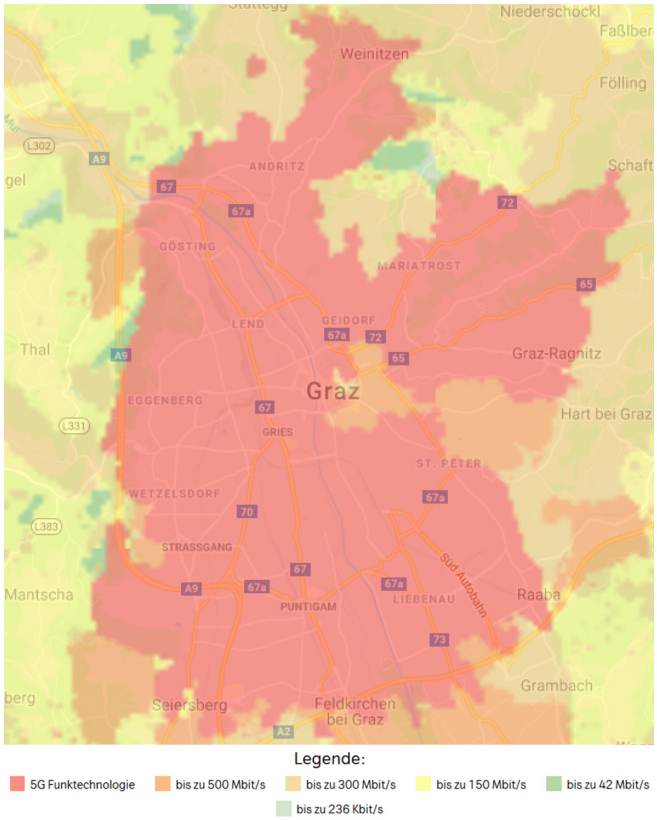


Fig. 2 5G mobile network coverage in Graz, Austria [4]

In the regard of bandwidth requirements, there are already development trends in the direction of the successor technologies 6G and 7G, see [5] and [6]. The density and thus also the power flow densities of the various mobile radio transmission systems, which affect the individuals, are linked to the increase in bandwidth requirements. It should not be overlooked that there are many other high-frequency wireless services in addition to mobile radio technologies, such as (digital) radio (DAB, DVB-T2, etc.), but also wireless networks in commercial and home use (WLAN). In the context of this publication, long-term measurements of the power flux densities of the different radio and mobile radio technologies in the inner-city area of Graz, Austria, which is completely covered by 5G (check Fig. 2) are presented and compared. Furthermore, the results are discussed in relation to health aspects [7], regulatory and legal norms. The increased use of mobile networks is particularly evident in times of home quarantine, due to the high number of home offices caused by the COVID-19 crisis which is also reflected in higher power flux densities. In the light of the current

circulating unsubstantiated and incomprehensible conspiracy theories and rumors regarding 5G, especially in social media, this publication is also intended to provide a comparative neutral basis of understanding based on real long term measurements with state of the art equipment for basic technical and physical understanding.

II. Law- and Health Issues

Specific exposure limit values (ELVs) and action levels by the directive 2013/35/EU of the European Parliament and the council 26 June 2013 on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (ALs) especially regarding electromagnetic fields and non-thermal effects, as well as health aspects have been mentioned in an earlier publication [7]. The relevant reference levels of exposure did not change so far and are outlined again in TABLE I.

TABLE I. Reference levels for electric, magnetic and electromagnetic fields [8]

Frequency range	Reference levels			
	E-Field (V/m)	*H-Field (A/m)*	*B-Field (μT)*	S_{eq}[a] *(W/m²)*
400 MHz — 2 GHz	$1.375 \cdot \sqrt{f}$ [b]	$0.0037 \cdot \sqrt{f}$ [b]	$0.0046 \cdot \sqrt{f}$ [b]	$\frac{f}{200}$
2 — 300 GHz	61	0.16	0.20	10

[a] Equivalent plane wave power density

[b] f as MHz

Furthermore, it should be noted that the risk of a tumour not only increases with the intensity of the radiation exposure, but also the accumulation of the exposures favours the probability getting a tumour [9].

III. Test and Measurement Setup

The electromagnetic radiation caused by different sources, like a DVB-T2 transmitting station in an inner-city area (distance around 2 km), a state of the art WLAN access point (2.4 and 5 GHz) in a distance of around 2 meters in an office, the mobile phone and 5G frequency bands in an Austrian inner-city area have been measured over a timespan of a number of days during the time of COVID-19 exit restriction in April 2020. The power spectrum and the power density at the different frequency bands have been recorded for around 24 hours.

Fig. 3 Measurement setup in the office

The measurement was done with a calibrated spectrum analyser and a PC to record the measurement data (Fig. 3).

IV. Measurement Results

In this chapter the measurement results regarding the different radiation sources are presented, explained and finally compared to the legal limits. First the television standard DVB-T2 is dealt with, then the two WLAN frequencies (2.4 GHz and 5 GHz) and finally the mobile phone network spectrums 2G, 3G, 4G and 5G.

A. Measurement results regarding DVB-T2

The power flux density depending on frequency caused by DVB-T2 is presented in Fig. 4. The measurement period here was 24 hours. The maximum of $S_{max,DVB\text{-}T2} = 42.630\ \mu W/m^2$ can be seen at 0.512 GHz.

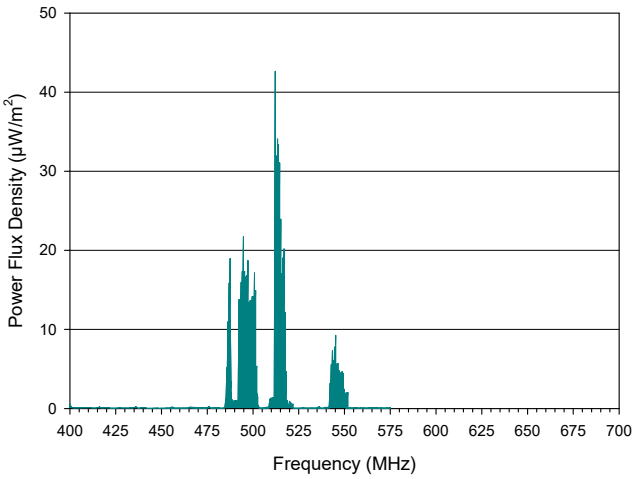


Fig. 4 Maximum power flux density of DVB-T2 versus frequency recorded over 24 hours

The magnetic field strength H can be calculated using

$$H = \sqrt{\frac{S}{Z_0}}, \tag{1}$$

where S is the measured power flux density and Z_0 the characteristic impedance of vacuum. Let there be $Z_0 = 120\pi\Omega \approx 377\ \Omega$ and $S_{max,DVB\text{-}T2} = 42.630\ \mu W/m^2$, the maximum magnetic field strength comes to $H_{max,DVB\text{-}T2} = 0.336$ mA/m. By substituting the calculated value of $H_{max,DVB\text{-}T2}$ into

$$E = Z_0 \cdot H, \tag{2}$$

the maximum electric field strength reaches $E_{max,DVB\text{-}T2} = 126{,}773$ mV/m. Finally, the magnetic flux density B is obtained using

$$B = \mu_0 \cdot \mu_r \cdot H, \tag{3}$$

where $\mu_0 = 1.257\ N/A^2$ is the vacuum permeability and $\mu_r = 1$ the relative permeability for air. By inserting μ_0 and μ_r, the magnetic flux density comes to $B_{max,DVB\text{-}T2} = 0.423$ nT. Inserting the frequency f = 0.512 GHz into TABLE I the limits of the magnetic and electric field are $H_{lim,0.512GHz} = 83.721$ mA/m and $E_{lim,0.512GHz} = 31113$ mV/m and the limit of the magnetic flux density is

$B_{lim,0.512GHz}$ = 104.086 nT. All maximum values fall within the limits.

B. Measurements results regarding WLAN

Both the 2.4 GHz and the 5 GHz band were measured and analysed.

1) 2.4 GHz WLAN band

In Fig. 5 the maximum power flux density of 2.4 GHz WLAN depending on the spectrum is shown. It was measured for 24 hours. A power flux density of $S_{max,2.4GHzWLAN}$ = 21.110 µW/m² was reached at f = 2.418 GHz.

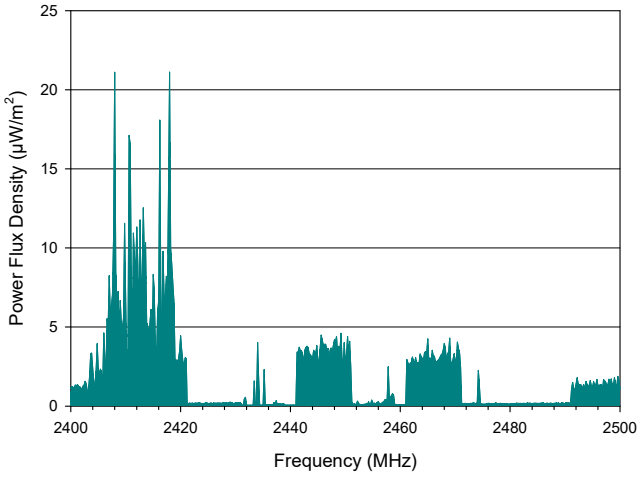


Fig. 5 Maximum power flux density of 2.4 GHz WLAN versus frequency recorded over 24 hours

By means of (1) and the measured value $S_{max,2.4GHzWLAN}$, the maximum magnetic field strength caused by the mobile device is $H_{max,2.4GHzWLAN}$ = 0.237 mA/m. $H_{max,2.4GHzWLAN}$ in (2), the calculated electric field strength reaches $E_{max,2.4GHzWLAN}$ = 89.210 mV/m. Using $H_{max,2.4GHzWLAN}$ in (3) leads to the magnetic flux density of $B_{max,2.4GHzWLAN}$ = 0.297 nT. The legal limits of the magnetic and electric field at the frequency of f = 2.418 GHz are $H_{lim,2.418GHz}$ = 160 mA/m and $E_{lim,1GHz}$ = 61000 mV/m. The magnetic flux density is limited to $B_{lim,2.418GHz}$ = 200 nT. Again all maximum values fall within the limits.

2) 5 GHz WLAN band

Fig. 6 presents the power flux density of 5 GHz WLAN depending on the spectrum. The measurement lasted 24 hours. The maximum power density of $S_{max,5GHzWLAN}$ = 88.950 µW/m² was reached at f = 5.711 GHz.

Note: The blanking in Fig. 6 from around 5.550 MHz to around 5.660 MHz is an access point hardware and regulation measure to reduce interferences regarding flight, weather and other types of radar working in this frequency band and to enable reasonable coexistence of both systems.

By means of (1), (2), (3) and $S_{max,5GHzWLAN}$ the following values were calculated. The maximum magnetic field strength is $H_{max,5GHzWLAN}$ = 0.486 mA/m, the maximal electric field strength $E_{max,5GHzWLAN}$ = 183.123 mV/m and the magnetic flux density $B_{max,5GHzWLAN}$ = 0.610 nT. The same legal limits apply here as for 2.4 GHz, see TABLE I. Again all maximum values fall within the limits.

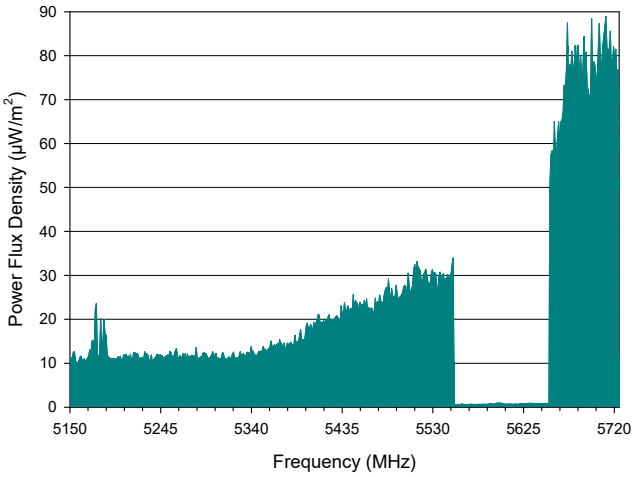


Fig. 6 Maximum power flux density of 5GHz WLAN versus frequency recorded over 24 hours

C. Measurement results regarding mobile phone network spectrums

First 5G spectrum is analysed followed by the other once in a separate subchapter.

1) 5G mobile phone network spectrum

The next results are for 5G. The power flux density depending on frequency was measured for 24 hours (compare Fig. 7). The maximum power density $S_{max,5G}$ = 46.720 nW/m² is located at 3.758 GHz. This maximum value is about three powers of ten lower than the values measured so far.

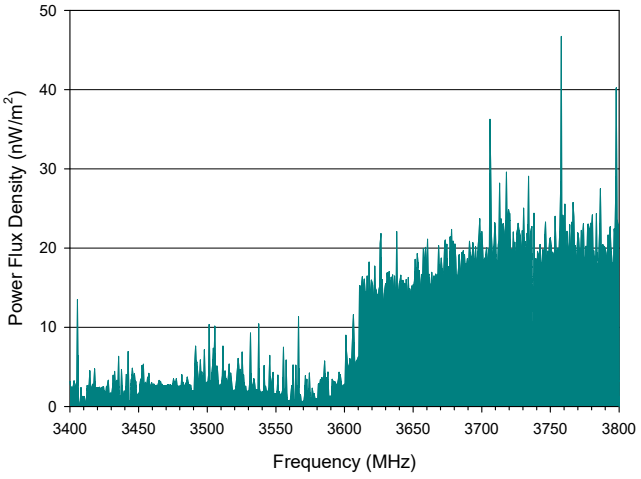


Fig. 7 Maximum power flux density of 5G mobile phone network versus frequency recorded over 24 hours

Using equations (1), (2), (3) and $S_{max,5G}$, the following values can be calculated. The maximum magnetic field strength is $H_{max,5G}$ = 0.011 mA/m, the maximal electric field strength $E_{max,5G}$ = 4.197 mV/m and the magnetic flux density $B_{max,5G}$ = 0.014 nT. The same legal limits apply here as for 2.4 GHz, because the frequency range is above 2 GHz (see TABLE I). Again all maximum values fall within the limits.

2) 2G, 3G and 4G mobile phone network spectrums

And last but not least, 2G, 3G and 4G are covered here. It was measured for 24 hours. The maximum power density $S_{max,2G3G4G} = 175.9\ \mu W/m^2$ at 0.927 GHz can be clearly seen in Fig. 8.

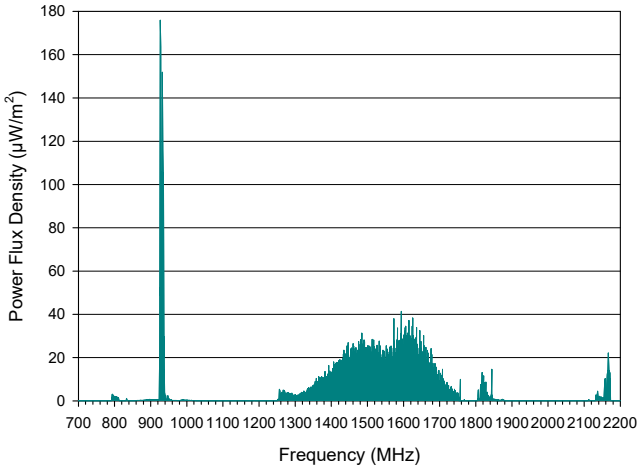


Fig. 8 Maximum power flux density of 2G, 3G and 4G mobile phone network versus frequency recorded over 24 hours

Assisted by (1), (2), (3) and $S_{max,2G3G4G}$, the following values were calculated. The maximum magnetic field strength is $H_{max,2G3G4G} = 0.683$ mA/m, the maximal electric field strength $E_{max,2G3G4G} = 257.516$ mV/m and the magnetic flux density $B_{max,2G3G4G} = 0.858$ nT. The following values were calculated using TABLE I and the frequency f = 0.927 GHz for the legal limits. The magnetic field is limited to $H_{lim,0.927GHz} = 112.653$ mA/m, the electric field to $E_{lim,0.512GHz} = 31113$ mV/m and the magnetic flux density to $B_{lim,0.512GHz} = 140.055$ nT. Again all maximum values fall within the limits.

D. Summary of the results

In TABLE II, TABLE III and TABLE IV the measured and calculated results are compared to the legal limits. All measured and calculated results are within the legal exposure limits.

TABLE II. SUMMARY OF RESULTS AND LIMITS OF DVB-T2

	DVB-T2 Fields		
	***E* (V/m)**	***H* (mA/m)**	***B* (nT)**
Legal Limits	31.11	83.72	104.09
Measurement	0.127	0.336	0.423

TABLE III. SUMMARY OF RESULTS AND LIMITS OF WLAN

	2.4 GHz WLAN band			5 GHz WLAN band		
	***E*-Field (V/m)**	***H*-Field (mA/m)**	***B*-Field (nT)**	***E*-Field (V/m)**	***H*-Field (mA/m)**	***B*-Field (nT)**
Legal Limits	61.00	160.00	200.00	61.00	160.00	200.00
Measure-ment	0.089	0.237	0.297	0.183	0.486	0.610

TABLE IV. SUMMARY OF RESULTS AND LIMITS OF MOBILE PHONE NETWORK SPECTRUMS

	5G spectrum			2G, 3G, 4G spectrum		
	***E*-Field (V/m)**	***H*-Field (mA/m)**	***B*-Field (nT)**	***E*-Field (V/m)**	***H*-Field (mA/m)**	***B*-Field (nT)**
Legal Limits	61.00	160.00	200.00	41.86	112.65	140.06
Measure-ment	0.004	0.011	0.014	0.258	0.683	0.858

V. CONCLUSIONS

The measurement data clearly has shown that the power flux densities of the different radio and mobile radio technologies in an inner-city area of Graz, Austria are below the legal limits. In [7] the maximum radiation caused by the mobile phone was measured, $S_{max,phone} = 217\ mW/m^2$ was reached at around f = 1 GHz.

The highest power flux densities were measured in a 2G, 3G and 4G mobile phone download link band with $S_{max,2G3G4G} = 175.9\ \mu W/m^2$ at a frequency of 926.5 MHz. Also to mention is the measurement result of a TV broadcast station in around 2 km distance with $S_{max,DVB-T2} = 42.630\ \mu W/m^2$ at a frequency of 512.2 MHz compared to the measurement results regarding a standard 2.4 GHz and 5 GHz WLAN Access point in a distance of around 2 meters to the measurement device with $S_{max,5GHzWLAN} = 88.950\ \mu W/m^2$ at a frequency of 5.711 GHz.

The DVB-T2 broadcast station Graz 9 in the center of Graz transmits with a maximum effective radiated power of $ERP_{max} = 5.623$ kW[10].

For comparison, in [7] the maximum radiation caused by a mobile phone device was measured, $S_{max,phone} = 217\ mW/m^2$ was reached at around f = 1 GHz. The corresponding magnetic and electric field strength did not exceed the legal limits; the values came to one fifth of the limits. However, the specific energy absorption rate (SAR) was exceeded by approximately two times when using the device directly on the head.

Surprisingly, the power flux densities in the 5G spectrum range in the measurement period were very low (in the nW/m^2 range). Especially since the use of 5G is just in the beginning this result is just a momentary statement and is also comprehensible regarding the current low spread of 5G end devices. It will therefore be necessary to measure the radiation exposure of mobile radio technologies in the future on a regular basis in order to ensure the legal limits and to reduce possible health hazards. It is also emphasized that the radiation power of a mobile phone at the ear exceeds the SAR limit by a factor of 2, and that even a short distance extension between the end device and the skull can achieve a significant reduction in exposure and thus a reduction in any health impairments.

[1] P. Jonsson *et al.*, 'Ericsson Mobility Report. November 2019', Stockholm, Sweden, 2019. Accessed: Apr. 15, 2020. [Online]. Available: www.ericsson.com/mobility-report.

[2] K. Rixecker, 'Youtube entfernt Verschwörungsvideos, die 5G mit Corona in Verbindung bringen', Apr. 09, 2020. https://t3n.de/news/youtube-entfernt-5g-corona-1269176/ (accessed Apr. 15, 2020).

[3] P. Dax, 'Corona-Panik: Dutzende 5G-Masten in Europa angezündet | futurezone.at', 2020. https://futurezone.at/digital-life/corona-panik-dutzende-5g-masten-in-europa-angezuendet/400815629 (accessed Apr. 18, 2020).

[4] A1 Telekom Austria AG, ‘Netzabdeckungskarte: Start’. https://www.a1.net/hilfe-support/netzabdeckung/frontend/main.html (accessed Apr. 15, 2020).

[5] M. Latva-Aho and K. Leppänen, *Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence - 6G Research Visions 1, September 2019*, no. September. Oulu, Finnland, 2019.

[6] R. Saracco, ‘What about 7G? – IEEE Future Directions’, 2019. https://cmte.ieee.org/futuredirections/2019/03/23/what-about-7g/ (accessed Apr. 18, 2020).

[7] P. Mandl, P. Pezzei, and E. Leitgeb, ‘Selected Health and Law Issues Regarding Mobile Communications with Respect to 5G’, in *Proceedings - 2018 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications, CoBCom 2018*, Aug. 2018, doi: 10.1109/COBCOM.2018.8443980.

[8] CEU, ‘Council Recommendation of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields (0 Hz to 300 GHz)’, *Off. J. Eur. Communities*, vol. L 199, pp. 59–70, 1999, Accessed: Mar. 14, 2017. [Online]. Available: https://www.bmvit.gv.at/telekommunikation/recht/europa/empfehlungen/downloads/em1999en519.pdf.

[9] F. Pareja-Peña, A. M. Burgos-Molina, F. Sendra-Portero, and M. J. Ruiz-Gómez, ‘Evidences of the (400 MHz–3 GHz) radiofrequency electromagnetic field influence on brain tumor induction’, *International Journal of Environmental Health Research*. Taylor and Francis Ltd., 2020, doi: 10.1080/09603123.2020.1738352.

[10] Forum Mobilkommunikation, ‘sendekataster.at’. https://www.senderkataster.at/karte (accessed Apr. 16, 2020).

An Iteratively-Improving Internet-of-Things Honeypot Experiment

Urban Sedlar, Leon Štefanič Južnič, Mojca Volk
Faculty of Electrical Engineering, University of Ljubljana, Trzaska cesta 25, SI-1000, Ljubljana, Slovenia
urban.sedlar@fe.uni-lj.si, leon.stefanic@ltfe.org, mojca.volk@fe.uni-lj.si

Abstract—In this paper we present a prototype implementation of an iteratively improving low-interaction Internet-of-Things (IoT) honeypot, based on serving responses of real IoT devices obtained through IoT search engines, as well as devices and services under our own control. The experiment was designed to confirm if this is a viable approach to mimicking a heterogeneous group of black-box devices. In the experiment we focused on only one of the protocols used in the IoT world, the Hypertext Transfer Protocol (HTTP), primarily due to widespread use and mature tooling. Our findings show that it is trivial to learn enough responses to induce deeper probing, and that some of the knowledge discovered in this way could not have been obtained by using any other publicly available resources.

Keywords—cybersecurity, internet of things, IoT, honeypot

I. INTRODUCTION

In the past decade, we have witnessed unprecedented proliferation of the Internet of Things (IoT). This emerging concept has revolutionized societies and industries, and today presents extremely promising technology with good business forecasts. According to IDC and Statista, the number of connected IoT devices has already exceeded 30 billion in 2020 and the forecasts estimate a steady growth, resulting in projections for 2025 ranging anywhere between 40 and 75 billion connected IoT devices. As the technology moves beyond the hype and massive IoT becomes a reality with the imminent rollout of the Fifth Generation (5G) of mobile networks, a wide variety of new and exciting opportunities will emerge, such as autonomous vehicle communication, drone-assisted applications, telemedicine, extended reality and many more.

However, such explosive growth of IoT has a dark side – with margins racing to the bottom, IoT devices often lack in security features and secure default settings. In addition, several studies have found a lack of long-term support and software upgrades across the industry [1][2], thus leading to large deployments of IoT abandonware: publicly exposed poorly protected devices, many of which can directly affect safety, privacy and security of citizens.

These threats are not just theoretical. Several organizations and initiatives monitor the prevalence of such cyberattacks: F-Secure's attack monitoring statistics for 2019 and SonicWall's 2019 Cyber Threat Report confirm a concerning increase in the volume of detected IoT attack attempts, with a detected 55 % increase in IoT malware attacks in 2019 compared to 2018.

In recent years, another very promising field has emerged in relation to IoT security – the distributed ledger technology (DLT). DLT comprises an immutable ledger of records, where mutual trust, data integrity and in some cases, confidentiality, can be enforced by strong cryptography. DLT can either provide an IoT data exchange layer or serve as a decentralized platform to deploy smart contracts – programs that can control anything from device ownership to device functionality [3][4]. For that reason, we see it as an important part of the future IoT landscape and have included it in the analysis. It is worth noting that DLT nodes and devices are an extremely high-profile target themselves, since the reward for compromising such a system can often be purely financial (i.e., stolen cryptocurrency) and the volume of cyber-attacks is further expected to escalate.

In such landscape, active targets, such as honeypots, are instrumental for assessing the threat level and modus operandi of either targeted or untargeted attacks and hacking attempts, and providing valuable information for defense and rapid incident response. In IoT, however, vulnerabilities are typically highly dependent on specific device brand or even firmware version, which forces the attackers to perform several checks to gather more device info prior to launching the attack. In addition, the IoT landscape has a highly scattered long tail of deployed devices, which makes it harder to specialize for and mimic any specific device. Consequently, designing an effective IoT honeypot with capabilities to respond to a growing level of heterogeneity in terms of IoT targets and attack variants continues to be a challenge of major interest. The most promising approaches are incorporating learning capabilities into an IoT honeypot to be able to adapt to a continuously morphing collection of new IoT attack variants and to mimic a varied range of IoT devices or services.

In this paper, we present an experiment evaluating the feasibility and efficacy of iteratively refining honeypot's responses. Section II presents background and the state of the art, Section III outlines our architecture, Section IV describes the implementation and Section V the results.

II. BACKGROUND

IoT devices communicate over a variety of protocols, acting as either clients or servers [5]. Properly implemented devices use client mode with protocols such as MQTT, HTTP, CoAP, etc., and only use outbound connections (e.g., to a MQTT broker or a HTTP server), while no ports are open on the device itself. Two issues with this are that (1) not all use cases and payload types are suitable for client-initiated communication (e.g., keeping alive a communication over a constrained network link to receive data) and (2) the correct server end-point needs to be configured on the device beforehand, for which some kind of server process on the device itself is often used. The latter is especially problematic

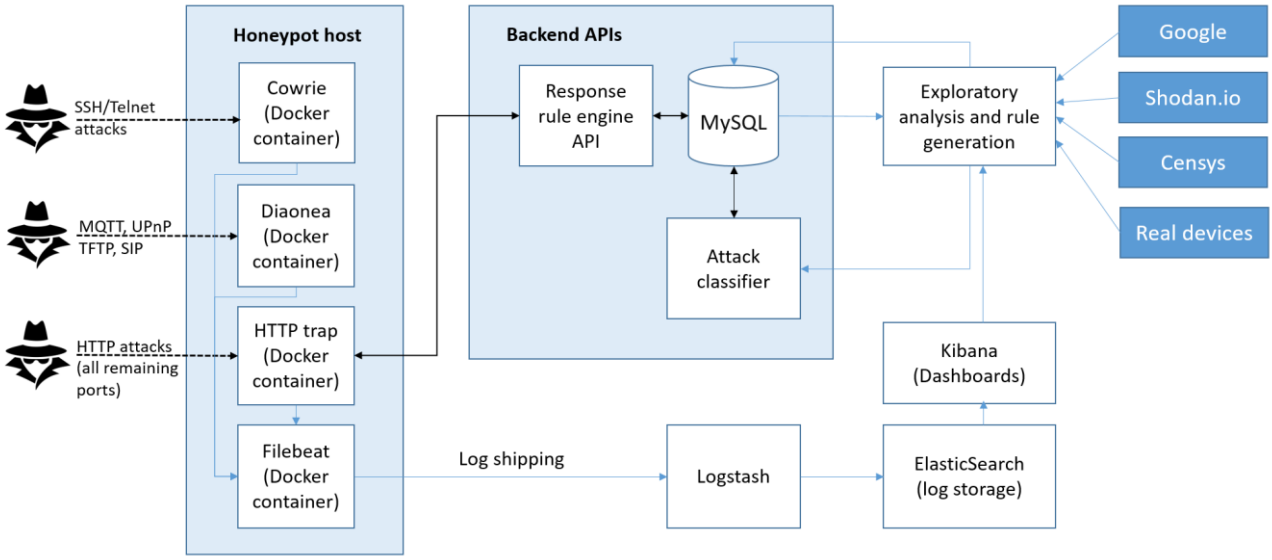


Figure 1: System architecture

with standalone (non-cloud-connected) consumer IoT devices using Wi-Fi, such as cameras, printers, lightbulbs, thermostats, etc. Exposing such devices to the public Internet thus presents a significant security risk, however it also adds value to many devices (e.g., remotely accessing a camera or a thermostat). This increased attack surface subsequently also presents a possible doorway to the private networks of the users, where many more unprotected devices can reside.

Most reviewed literature on the topic of IoT honeypots focuses on either end devices in server mode or the supporting server infrastructure, using any of the common communication protocols. This is primarily due to the large attack surface and remote exploitability, as opposed to physically manipulating a device to extract keys, install custom software, etc. [6] Setting up a honeypot thus includes mimicking any of the following.

1) HTTP server on the device itself, which also includes all higher-level standards and protocols, such as SOAP, TR-069, and similar. This is common for cameras, digital video recorders, printers, routers, modems and other network and terminal equipment.

2) Supporting HTTP server backend infrastructure; this is today quite mature, as it is commonly used as part of web and mobile applications; exploits here thus have to be much more involved, and this option is not common for IoT honeypots.

3) MQTT broker infrastructure; since MQTT broker is not meant to be used on end devices, only broker infrastructure can be compromised to capture data or manipulate the data sent to the IoT devices. This presents a good opportunity for setting up a honeypot, especially since many MQTT broker deployments are unsecured out-of-the-box.

4) Constrained Application Protocol (CoAP) server infrastructure; CoAP is a lightweight request-response protocol based on UDP; similar considerations apply as with MQTT.

5) Extensible Messaging and Presence Protocol (XMPP) server infrastructure; XMPP was initially used for instant messaging, but has had a reneissance as an IoT protocol. similar considerations apply as with MQTT and CoAP.

6) Telnet and SSH servers on the device itself; this is extremely common in more powerful devices running a full-fledged operating system (e.g., Linux). Gaining access through telnet and SSH can often result in privilege escalation, which allows the attacker full control of the device.

7) FTP and TFTP, typically as a part of infrastructure; some devices, such as cameras, might use a FTP server for download of the content; however, the scope of the damage that can be done is typically very limited.

8) Universal Plug and Play (UPnP), which is usually exposed on a local LAN and due to broadcast packets inaccessible from public internet. However, UPnP is responsible for a significant part of exposed devices, since it allows devices on local networks to set up port forwarding without user's knowledge.

9) other less popular protocols.

There are several generic and IoT-specific software packages that specialize in the listed protocols [7], such as Cowrie (Telnet/SSH), Dionaea (HTTP, MQTT, FTP, TFTP, UPnP), HoneyPy (CoAP, TFTP, TR-069) and TelnetIoT (Telnet); most of the published research focuses on analysis of the data from these low interaction honeypots [8][9][10]. However, a significant problem with low interaction honeypots is their persuasiveness – they typically only implement generic prepared responses or simple state machines that fail at deeper probing, causing the attacker to lose interest, or the attack script to fail a prerequisite check.

To solve these problems, important research is going on to extend these low interaction approaches into more convincing systems. IoTPOT [11] is an example of a reactive honeypot that emulates Telnet service in a variety of IoT devices; its key feature is dynamically generated responses created based on emulation of real IoT devices in a virtual environment. IoTCandyJar [12] is another example and represents an adaptive honeypot using machine-learning techniques to determine the most appropriate tactics to prolong an attack; it makes use of publicly available scanning tools and vulnerability crawlers (Shodan, Censys, Masscan and Zoomeye) to find and then probe real-world IoT devices, collect their response and using machine learning trains its response tactic. Honware [13] is a similarly general approach that uses advanced techniques to emulate various real-world firmware images, which gives it high persuasiveness, but involves a lot of manual work for image preparation. Unfortunately, no source code is available for these research efforts, however we can learn a lot from the methodology and test the viability of such concepts in our own efforts, confirming if iteratively improving IoT honeypot present a good opportunity for further study.

III. System architecture

The architecture of the experiment is shown in Figure 1. The honeypot itself consists of an event-driven HTTP server that listens on a very wide port range – claiming over 60k of the possible 65535 TCP ports. However, the response to the requests is not completed immediately; instead, the request data (source and destination IP address, source and destination TCP port number, HTTP verb and all HTTP headers) are sent to a central backend API, which first stores all request parameters and then decides how to respond. The response is determined by a rule engine, taking into account the request parameters and previous context.

All requests are logged in the backend relational database and written to a local log file, which is shipped to a central logging facility. In the big picture, the HTTP server results are complemented with logs from a SSH/Telnet honeypot and Dionaea, an IoT honeypot that listens on a limited number of ports. All three honeypots are containerized and can be deployed in concert on any host, or managed using orchestrators such as Docker Swarm or Kubernetes. The stateless nature of the honeypot node makes it easy to scale the system to a large number of hosts.

Having a central backend and database for providing responses is suboptimal, as it adds a delay of one round-trip time plus the request processing time. Careful timing measurements, as have already been described in the scientific literature [14], could immediately reveal that the responder is not a real IoT device; however, with the first version we are discounting such sophisticated attacks and expect to target mostly script kiddies. There is however much room for improvement by caching the entire database with rules and locating the response logic on the honeypot node itself.

The classifier script running on the backend periodically tags requests with appropriate labels based on simple heuristics. This is crucial to aid the operator in determining which attacks are targeting classical server infrastructure, and which are IoT-specific, making it easier to explore IoT attacks in more detail.

IV. Implementation

A. Data collection

Data collection component was built in two iterations; in both the web server was developed in Node.js. We used multiple instances of the built-in and lightweight *http* server (via *http.createServer*), one per port. In the first phase of the experiment, we wanted to establish a baseline dataset, for which we covered almost all of the 65535 ports. In the second phase, a more limited list of ports was used.

The Node.js server handles the incoming request and pushes all the data to a back-end API, where it is stored in a MySQL database table. Another table is used as the response generation rule set, listing response templates against regex conditions for the TCP port, URL path, HTTP verb, request headers, and request body. For each rule, a HTTP status code, response headers object, and response body are provided in the form of a template, where dynamic placeholders can be substituted (for example, random strings, random numbers, timestamps, etc.). Each rule also has a priority, which can be used to promote or demote it, and special flags to deactivate it, turn off all logging, return random binary payload of specific length, etc. In addition to logging the data to the backend directly upon each request, the data is also written to a local log file, which is shipped to an independent ElasticSearch cluster using Filebeat and Logstash services, and visualized using the Kibana software package. This so-called ELK stack presents a robust log-shipping pipeline and can be reused across multiple honeypots. In the second phase of the experiment we deployed two additional honeypots: a Cowrie SSH/Telnet honeypot and Dionaea in parallel, reusing the same log-shipping stack. In the first phase, due to the requirement of as many available TCP ports as possible, we deployed the collector node on a dedicated server. In the second phase, due to reduced number of ports, we were able to deploy it as a Docker container; which we manage with Kubernetes.

B. Classification

Next, to find out which requests target IoT devices, we developed a simple rule-based classifier. We used an iterative approach, where our main strategies and methods were:

- Exploratory analysis and device endpoint lookups in search engines (generic web search engines such as Google and DuckDuckGo, as well as more IoT-device specific Shodan.io and Internet service-specific Censys.io);
- Excluding well-known cloud-based service endpoints (some common examples are big data processing Hadoop framework, various infrastructure monitoring frameworks, Oracle WebLogic and other application servers, Joomla and WordPress content management systems;
- Excluding white hat web crawlers and analyzers that can be identified by their user agent string (e.g., GoogleBot), as well as by their access the robots.txt;
- Additionally excluding heavyweight web development frameworks and languages that are uncommon in the world of low-power IoT devices (Java Server Pages, PHP, ASP).

V. RESULTS

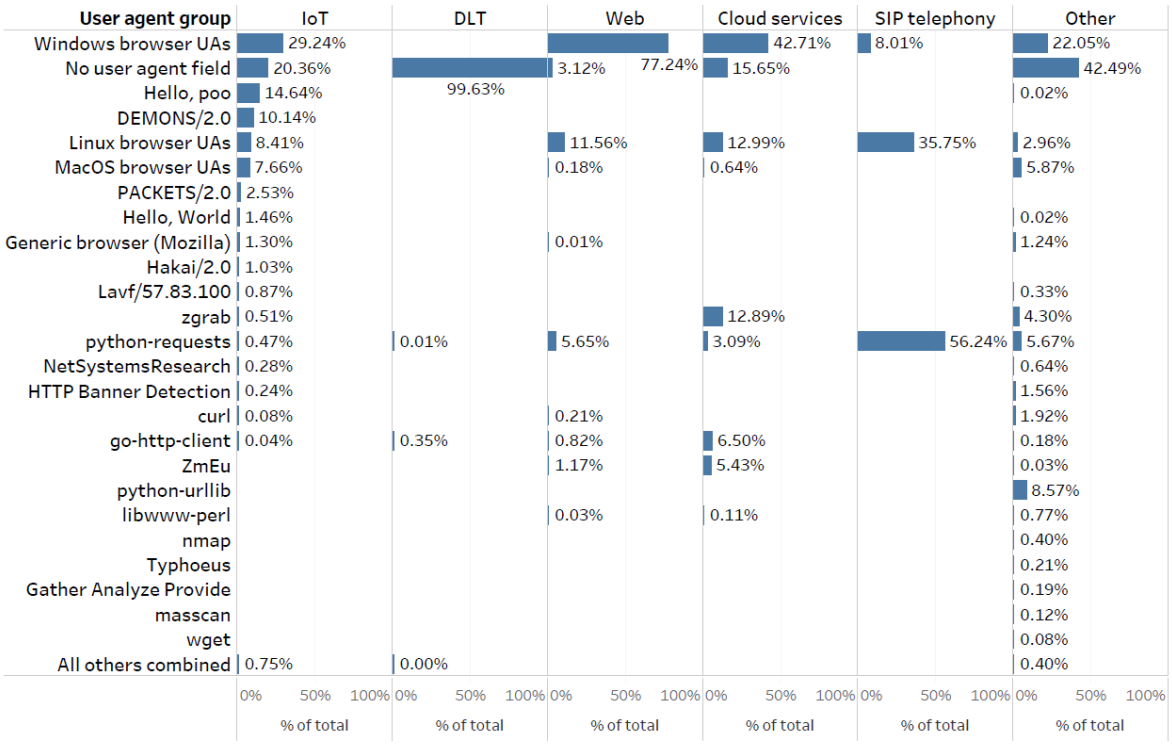


Figure 2: User agent vs. target classification in the baseline dataset.

In the period of 30 days in January and February 2020 we captured over 70k baseline HTTP requests over the entire range of TCP ports (65535 ports with 2 deliberate high-traffic non-HTTP exceptions: 23—Telnet and 53—DNS). All requests returned a 200 OK empty response. The collection process was done on a single node with no other services running, with the goal of providing information about the port popularity and to inform us where to focus our efforts. The distribution of the ports with valid HTTP requests is shown in Figure 3.

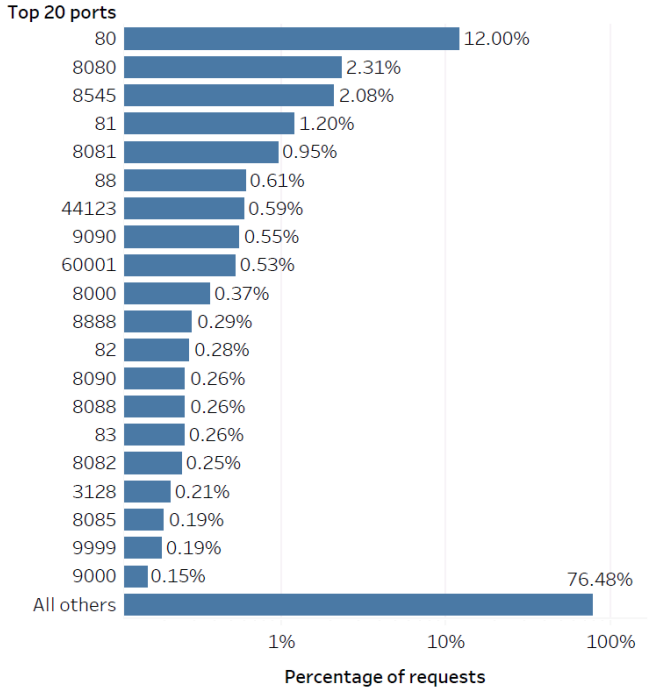


Figure 3: Top 20 TCP ports by the share of valid HTTP requests

Based on this dataset, we proceeded with classification of requests according to the described heuristics. A large number of requests provided no tangible information about either the source (as would be, for example, request headers indicating a vulnerability scanner) or the intended destination (e.g., a URL path or a post request body, indicating the intended target). Thus, there is a large residual group "Others". The breakdown obtained through classification is shown in TABLE I. Based on our heuristic, of N=70.659 HTTP collected baseline requests only 1.46% were IoT-related.

In the second phase of the experiment, several popular devices (IP cameras and routers) on popular TCP ports were identified through Shodan.io and crawled for basic landing-page and unauthorized login response. These responses were captured and converted into response rules. The same was performed for blockchain nodes; however, we used publicly available documentation and our own server instead of third-party nodes to learn responses.

TABLE I. BASELINE REQUEST CLASSIFICATION

Classification	Percentage of requests
Cryptocurrency APIs (Ethereum)	16.22%
Web: CMS, heavyweight frameworks, crawlers	9.62%
IoT: cameras, DVRs, routers, modems	1.46%
SIP telephony: Asterisk, FreePBX, Elastix	0.69%
Cloud infrastructure: databases, Hadoop, monitoring infrastructure	0.94%
Others (ambiguous or unidentified)	71.07%

The experiment was also scaled to 5 nodes in the second phase, while the number of concurrent HTTP servers was limited to the 50 most popular ports, which was done in preparation of a larger scale deployment in managed environments.

A. *Case study 1: an internet-connected camera*

Our most successful IoT attempt was an internet-connected camera; we have identified a popular example on Shodan.io, using the determined top ports as a constraint. Next, we crawled the landing page, fetching data from all included URLs, as well as the login button action. All response headers and page bodies were saved, allowing us to serve identical response as the original device. For example, the *IPCamera* landing page (depicted in Figure 4) has multiple JavaScript and CSS includes that were saved as well, and paired with exact match regular expressions shown in TABLE II.

Figure 4: Example of a camera landing page snapshot on TCP port 81

TABLE II. INITIAL RULE SET EXAMPLE

Port rule regex	HTTP verb rule regex	Path rule regex	Header and body rule regex
^81$	^GET$	^/$	.*
^81$	^GET$	^/ui.css$	.*
^81$	^GET$	^/lang_english.js$	.*
^81$	^GET$	^/lang.js$	.*
^81$	^GET$	^/tool.js$	.*
^81$	^GET$	^/devinfo.xml$	.*
^81$	^GET$	^/login.xml.*$	.*

After these initial "seed" responses, the ruleset was left in active state for two weeks, during which an increase in activity has been observed. The requests during this period fall into 2 categories:

- Requests that are indistinguishable from browser traffic of a user trying to load the camera page, then entering several username and password combinations, and finally leaving; search engines such as Shodan.io also fall into this category, but don't perform any login attempts; Shodan.io mimics a web browser when crawling and is detectable only by the source IP ranges.
- Requests of specialized scripts, which reveal much more – probing several additional endpoints that we were unable to discover on the real device.

In the latter case, for the mimicked *IPCamera* we obtained 164 different base URL endpoints (ignoring get query parameters). We inspected them and removed 8 with signs of exploits (get parameters including system commands). We crawled the remaining 156 endpoints, and of these:

- 93 returned a redirect to error page
- 55 returned redirects to login page, which means that was a valid endpoint
- 4 returned an entire login page body
- 4 returned an XML with an "unauthorized" message
- 1 endpoint returned just a string 233331; after the fact we could not find this behavior using any web search engines, nor was this string found in any GitHub repository, which means that the scanning software is not openly published. Thus, this is a perfect example of an obscure check to determine if the target is a real device or a honeypot.

By including the newly crawled responses, we have significantly extended the initial "seed" rule set, thus increasing the persuasiveness of the honeypot.

Figure 5 captures the user agents of the software probing the *IPCamera* in the second phase of the experiment. Only unique sessions are shown – all the subsequent requests are excluded (e.g., JavaScript and CSS files). This reveals some interesting insights into what software the attackers use to discover new resources. The high prevalence of Windows-based browser user agents could be related to the fact that the honeypot has been indexed by Shodan.io within the same day. (determined based on publicly available lists of Shodan.io scanner IP addresses), which has made it easily discoverable by anyone.

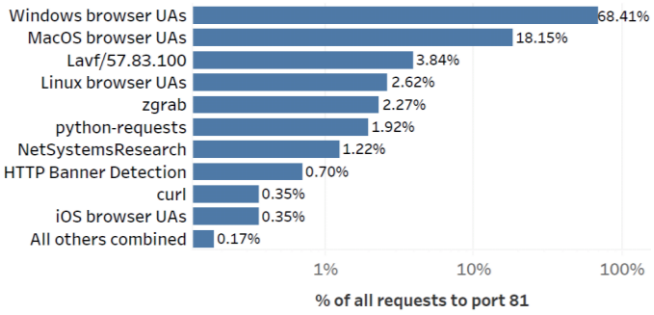


Figure 5: User agents making requests to the IPcamera honeypot

B. *Case study 2: Ethereum DLT node*

As already discussed, DLT networks and platforms present a promising solution to many trust and security-related IoT-problems. Ethereum is currently the most mature and rapidly evolving technology that unfortunately encompasses both a Smart Contract platform as well as a cryptocurrency. Thus, in this section we ponder if in this case the cure might be worse than the disease. The Ethereum node *geth* listens on TCP port 8545, and we see already in Figure 3 that this is the third most popular port by the number of requests.

Since the source code behind the API is open and actively developed, it is unlikely that an actual vulnerability is present there. However, a *geth* RPC port should never be publicly exposed, since anyone accessing it could also transfer the cryptocurrency from the accounts; this is done by first calling

eth_accounts method, querying its balance, and then initiating a transaction.

In our baseline dataset, the majority of requests to 8545 were HTTP post requests performing the following RPC call to detect if this is a *geth* node:

```
{"id":0,"jsonrpc":"2.0","method":"eth_blockNu
mber"}
```

After setting up the response rule matching the "jsonrpc" and "eth_blockNumber" strings, and returning random hex block identifiers, multiple calls followed, calling among others the following methods:

- eth_blockNumber
- eth_getBlockByNumber
- eth_accounts
- get_info
- web3_clientVersion
- net_version

These rules have caused a huge ramp-up of the requests to the port 8545, which in the second phase of the experiment accounts to 56% of all HTTP requests to the servers, up from 16% in the first phase.

An additional improvement to our approach would be to return the actual current block numbers instead of random strings.

VI. Conclusion and Future Work

We have demonstrated that IoT attacks are ubiquitous, and that proper responses can be iteratively refined to adequately convince attackers to reveal their knowledge and strategies. In the experiment, we focused on only one of the protocols used in the IoT world – HTTP – primarily due to widespread use and mature tooling; however in the future this could be extended to other protocols as well. The results show that with a real device serving as a model, our honeypot basically becomes an elaborate caching server with some added intelligence. Currently, the intelligence to modify responses has to be added through manual work; similarly, the learning process needs to be carefully supervised to avoid forwarding harmful requests to the end devices. Nonetheless, the approach shows great promise, and some of the discovered knowledge can be highly obscure, and could not be found using publicly available resources alone, without the help of such a crowdsourced approach.

Acknowledgment

This research was supported by the project FED4FIRE+, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 732638.

References

[1] Lin, Huichen, and Neil W. Bergmann. "IoT privacy and security challenges for smart home environments." Information 7, no. 3 (2016): 44.

[2] Neshenko, Nataliia, Elias Bou-Harb, Jorge Crichigno, Georges Kaddoum, and Nasir Ghani. "Demystifying IoT security: an exhaustive survey on IoT vulnerabilities and a first empirical look on internet-scale IoT exploitations." IEEE Communications Surveys & Tutorials 21, no. 3 (2019): 2702-2733.

[3] Pustišek, Matevž, Anton Umek, and Andrej Kos. "Approaching the Communication Constraints of Ethereum-Based Decentralized Applications." Sensors 19, no. 11 (2019): 2647.

[4] Khan, Minhaj Ahmad, and Khaled Salah. "IoT security: Review, blockchain solutions, and open challenges." Future Generation Computer Systems 82 (2018): 395-411.

[5] Karagiannis, Vasileios, Periklis Chatzimisios, Francisco Vazquez-Gallego, and Jesus Alonso-Zarate. "A survey on application layer protocols for the internet of things." Transaction on IoT and Cloud computing 3, no. 1 (2015): 11-17.

[6] Alladi, Tejasvi, Vinay Chamola, Biplab Sikdar, and Kim-Kwang Raymond Choo. "Consumer iot: Security vulnerability case studies and solutions." IEEE Consumer Electronics Magazine 9, no. 2 (2020): 17-25.

[7] Nawrocki, Marcin, Matthias Wählisch, Thomas C. Schmidt, Christian Keil, and Jochen Schönfelder. "A survey on honeypot software and data analysis." arXiv preprint arXiv:1608.06249 (2016).

[8] Sethia, Vasu, and A. Jeyasekar. "Malware Capturing and Analysis using Dionaea Honeypot." In 2019 International Carnahan Conference on Security Technology (ICCST), pp. 1-4. IEEE, 2019.

[9] Shrivastava, Rajesh Kumar, Bazila Bashir, and Chittaranjan Hota. "Attack detection and forensics using honeypot in IoT environment." In International Conference on Distributed Computing and Internet Technology, pp. 402-409. Springer, Cham, 2019.

[10] Banerjee, Mahesh, and S. D. Samantaray. "Network Traffic Analysis Based IoT Botnet Detection Using Honeynet Data Applying Classification Techniques." International Journal of Computer Science and Information Security (IJCSIS) 17, no. 8 (2019).

[11] Pa, Yin Minn Pa, Shogo Suzuki, Katsunari Yoshioka, Tsutomu Matsumoto, Takahiro Kasama, and Christian Rossow. "IoTPOT: A novel honeypot for revealing current IoT threats." Journal of Information Processing 24, no. 3 (2016): 522-533.

[12] Luo, Tongbo, Zhaoyan Xu, Xing Jin, Yanhui Jia, and Xin Ouyang. "Iotcandyjar: Towards an intelligent-interaction honeypot for iot devices." Black Hat (2017).

[13] Vetterl, Alexander, and Richard Clayton. "Honware: A virtual honeypot framework for capturing CPE and IoT zero days." In Symposium on Electronic Crime Research (eCrime). IEEE. 2019.

[14] Mukkamala, S., K. Yendrapalli, R. Basnet, M. K. Shankarapani, and A. H. Sung. "Detection of virtual environments and low interaction honeypots." In 2007 IEEE SMC Information Assurance and Security Workshop, pp. 92-98. IEEE, 2007.

Getting on Track – Simulation-aided Design of Wireless IoT Sensor Systems

Daniel Kraus
Pro2Future GmbH
Graz, Austria
daniel.kraus@pro2future.at

Konrad Diwold
Pro2Future GmbH
Graz, Austria
konrad.diwold@pro2future.at

Erich Leitgeb
Institute for Microwave and Photonic Engineering
Graz University of Technology
Graz, Austria
erich.leitgeb@tugraz.at

***Abstract*— When designing dependable communication systems for industrial application scenarios, the application environment where the solution will be integrated plays a crucial role. This study demonstrates how simulations can be used to integrate a target environment in the design process from an early stage. This allows to optimize wireless sensor networks (WSNs) in specific environments in terms of hardware requirements (e.g. antenna) and topology. The article discusses several application scenarios within the 2.4 GHz band and demonstrates how the design process can be aided by simulation to achieve dependable communication in harsh environments.**

Keywords—IoT, wireless sensor communication, network topologies, wireless communication protocols

I. INTRODUCTION

The vision of *the Internet of Things* (IoT) is to make communication ubiquitous and allow communication among as many "things" as possible. Ideally, all devices and objects should be connected and able to communicate without physical constraints. While wireless communication offers a potential solution for comprehensive connectivity, achieving reliable and dependable wireless sensor network communication in diverse industrial environments is a key challenge. There are plenty of scenarios and related research, where WSNs and IoT are applied and evaluated in industrial environments [1], [2]. In most applications, simple and cheap commercial off-the-shelf (COTS) devices are used to keep costs at a minimum. The communication underlying such applications often operates on the publicly accessible 2.4 GHz frequency band using protocols such as Bluetooth Low Energy (BLE), Zigbee, or other low-power protocols, as they allow for a long batterie life of the sensor nodes.

Unfortunately, COTS solutions often fail to provide dependable communication. Especially in environments, where metallic surfaces and electronics are abundant (e.g., industry, automotive), COTS-based communication devices deteriorate very quickly. There are many different approaches to improve wireless communication in such environments at the mentioned frequencies. Each environment poses an enormous amount of optimization potential when it comes to the devices, antennas, and position of the sensor nodes. Simulations can help to assess the quality of a potential solution in a target environment and optimize and tune the resulting communication system. Related research on this topic often focuses on finding fitting protocols for the adaptive selection of radio channels in harsh environments [3]-[5] and mainly covers the protocol layer, e.g., by realizing channel switching based on quality measurements. In most cases the radio signal strength indicator (RSSI) value is used to evaluate a communication channel, however, the results are still very dependent on the devices used and the RSSI value can be easily influenced by the devices' environment. Here a different approach is taken to evaluate the path loss and quality of the channel. Using detailed models and simulation allows to optimize the topology, antenna, and other communication parameters to minimize path loss.

The structure of the paper is as follows: In section (II), the key influences on wireless communications of harsh environments are discussed and analyzed. Based on the analysis, the most important communication parameters are discussed (III). Section (IV) outlines potential target environments and their simulation. Section (V) demonstrates the benefits of simulation in the context of devising a communication system for a traction vehicle. Finally, the paper is concluded and an outlook on future research is given (VI).

II. HARSH ENVIRONMENT

In the context of wireless communication, an environment is harsh, if many electronic devices, metallic surfaces, or other objects are present, which obstruct and interfere with the direct communication path. There has been extensive research on the physical impacts on WSNs [6], [7]. Various harsh communication scenarios have been evaluated for radio communications [8], [9]. Industry 4.0 scenarios are a good example, they provide a complex setting, where many components are moveable and thus, only wireless communication can be a reasonable solution for connectivity. Other examples are in- or inter-vehicle communication, which also poses a lot of inference and thus challenges towards wireless communication, but (as their industrial counterpart) requires dependable communication. The main differences between those scenarios are the distances and the mobility of wireless nodes. In e.g. a car, the position of the nodes will be fixed, and inter-node distances are between 2 to 4 meters. In an industrial environment, many nodes will be mobile, and distances can be easily up to a hundred meters.

There are many unique environments where a wireless communication link is still not reliable enough due to all unforeseen issues in communication. Minor changes in such constantly changing environments might affect the whole system's behavior. Consider transmitting devices interfering with other devices in their direct vicinity or frequency band congestions which can occur sporadically and could lead to a failure of communication. It is important to keep all these factors in mind when designing and optimizing wireless communication for harsh environments. Another direct impact on communication comes from effects such as varying temperatures, dirt, and vibrations, which degrade communication devices further [10]. Occurring physical effects in wireless communications are unique for various environments. The more complex an environment, the higher

the impact of such effects. A perfect line-of-sight (LOS) communication, where the signal can propagate along the most direct and unobstructed path, will rarely be achieved in a harsh environment. Therefore, many effects must be considered in the context of a wireless communication system design. The most important ones are described in (A)-(G). The described effects can be partly compensated if the right techniques and countermeasures are taken.

A. *Absorption*

All objects in the communication path act as an absorber. Metal surfaces are the worst in the direct vicinity of the transmitter because they absorb the most energy of a signal. If placed correctly, some surfaces can be also used to improve signal strength, if the transmitter is placed with the right angle and distance to this surface. The metal plane behind acts as a reflector and increases the signal strength. Other metallic surfaces in the vicinity will be still absorbing parts of the signal strength.

B. *Reflection*

Surfaces do not only absorb the signal they can also reflect signals. Any metallic surfaces next to the transmitter can lead to a reflection of the outgoing signal. The reflected signals hit the transmitter and reduce the signal strength of further outgoing signals.

C. *Scattering*

For short-distance communication, scattering might be the biggest issue. The transmitted signal hits a surface, edge, or other particles where the signal is scattered into different directions, causing a diffusion of the signal. The overall signal strength will be significantly decreased at this point of impact and the emerging secondary signals will be propagating into different directions. They might even arrive again as reflections at the receiver.

D. *Diffraction*

Similar to the scattering effect, the electromagnetic waves hit an edge or surface, where the signal is split or bent. With this effect, the signal can still get to the receiver, even if the communication path is directly blocked by an object. The overall signal strength of the diffracted signals is naturally much lower.

E. *Multipath*

If the direct path is blocked, two or more multipath signals will arrive at the receiver. They show a difference in signal strength and phase and will be recognized as interference at the receiving node. For short distances, this effect is weaker and can be often compensated. Destructive interferences do cause a fading of the signal, which might lead again to a very weak signal strength which might not be detected at the receiver.

F. *Interference*

Any devices or senders which are close to the wireless communication path and use the same frequencies for communication might be interfering with the signal.

G. *Doppler Effect*

Most of the investigated scenarios in the paper involve moving vehicles, thus the Doppler Effect must be considered. For in-vehicle communication, this effect is not decisive since all the nodes will travel at the same speed and the shift is everywhere the same. But when communicating with stationary nodes, there will be a significant shift in signals, depending on the difference in speed.

III. Identification of Communication Parameters

The environment with its physical constraints, components, and materials must be thoroughly investigated to eliminate as many factors as possible which might disrupt a wireless communication link. Additionally, a lot of time and costs can be saved if the wireless communication link is evaluated before testing it in the real environment.

Wireless communication in a harsh environment can take many forms. Adapting topology, protocol, hardware, and other aspects can improve a communication link. But also position and rotation of a transmitting device (especially the antenna) do have a lot of impacts. If any knowledge about the transmitting position is available, it is possible to adapt those parameters accordingly. Fig. 1 depicts a typical wireless communication link of a BLE node with its master node.

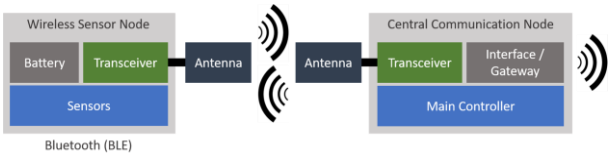


Fig. 1: BLE communication scenario between slave and master node

It requires a transceiver or at least a transmitter unit at each WSN node. Connected to this unit is an antenna to increase the transmitting strength of the outgoing signal. Signals and data are generated by electronics and sensors which are directly on the same circuitry or mounted and connected very closely to the node. The wireless sensor node and all its components are powered by a battery. Depending on how often and how long the device will be active, the battery must be chosen to fulfill the lifetime requirements.

The requirements of a wireless communication system vary for each application. Simple measurements with accelerometers, ultrasound, and other environmental (temperature, humidity, …) sensors require only a few kbit/s of data rate while more sophisticated sensors such as vibration sensors need at least a few hundred kbit/s. Wireless data transfer with vision-based sensors is not an option, only if the video stream is encoded at the transmitting device, which would require a power supply for the image processing unit.

In the following subsections, some parameters will be discussed which must be considered for a reliable communication link.

A. *Protocol*

The protocol layer is well-investigated and there are plenty of approaches and techniques for optimization. Some of them are mentioned in the Introduction. Almost all protocols are built on top of the well-investigated TCP or UDP transport layer protocols. BLE is the most established technology for low-power wireless networks and the newest version 5.x can achieve with certain settings application data rates of over 1 Mbit/s [11]. Other low-power wireless technologies achieve significantly lower data rates, which makes BLE the go-to technology for future applications, especially when it comes to IoT. The biggest issue with technologies such as BLE, Zigbee, etc. is that they all use the same 2.4 GHz frequency band, which is used by almost all other wireless communication techniques. Consequently, there are a lot of interferences on the physical level which can be only

compensated to some degree by the protocol. Devices should consume as little energy as possible and protocols have been the key element in the past, where most of the energy consumption could have been reduced. By the introduction of protocols such as Message Queuing Telemetry Transport (MQTT) and Constrained Application Protocol (CoAP), the overall protocol structure is reduced to the bare minimum for energy-efficient communication [12], [13]. In industrial environments, security can be a crucial factor to protect the systems' integrity. Using a security stack in the protocols will inevitably increase the power consumption of the devices. However, very promising work has been conducted on energy-efficient authentication protocols [14].

B. Topology

With an optimal layout of the network topology, many problems can be eliminated in advance. Therefore, it is of utmost importance to know the environment before choosing a topology. The strategy in many cases is that the structure should be relatively simple. Applications should utilize simple point-to-point or star network topologies. BLE and Wi-Fi are the best examples of this topology type. If the direct communication paths between two communication points are completely blocked, such a topology might easily fail. A better approach is then to use an adaptive topology, where the signal quality towards all neighboring nodes is used to route communication adaptively. This approach is much more sophisticated and data link, network, and transport layer must be optimized to fulfill the requirements for robustness and energy-efficiency. MQTT and CoAP apply multi-hop and meshed topologies to offer a highly dynamic behavior and network structure in industrial environments.

When only one central transceiver node is used (i.e., star topology), the positioning of this node in the environment will be challenging. It is almost impossible to find an "optimal" spot to have the best possible connection to all links at the same time since every node position might be unique with diverse characteristics. If the distance is large and a lot of interfering material is positioned in-between one node and another, it is often useful to use intermediate nodes that act as a signal amplifier. Then, the complexity on the other layers increases but the transmitting power of the transmitting node can be much lower. There are many possibilities for creating a specific topology for each unique environment. In a vehicle scenario, the sensor node positions will be fixed, while in industrial applications, many mobile components must be considered. Hence, a tradeoff has to be identified to achieve optimization regarding the topology.

C. Antenna

Another key element for communication, especially in harsh environments, is the antenna. The properties of antennas for wireless communications include small dimensions, minimal transmitting power, and omnidirectional radiation pattern (for movable objects). With a fitting antenna, optimized for certain positions, the quality of the communication link can be increased significantly. However, antenna design is a very time-consuming process and it is also unpracticable to use different designs for each position in an environment since the production of many devices with different antennas is not worthwhile. Still, an investigation of antenna types for specific environmental conditions seems to be a research area where a lot of potential for optimization is wasted. Antenna and device production are nowadays easier and cheaper than in the past, thus an approach based on individual antenna design for specific positions could be one key factor to achieve reliable wireless communication in harsh environments. In specific harsh environments, as discussed in section (IV), the choice of a fitting antenna with a more directed characteristic and an insensitivity to interferences yields much better results for designated positions. The transmitting power might be higher for certain antennas and will result in lower battery life. These decisions and drawbacks have to be considered carefully in the design process and the most important characteristics of the wireless link must be defined to estimate the key parameters for maximum output, reliability, and system availability. Investigations on parameter optimization of antennas have been conducted in previous work on this topic [15].

IV. Environments for Simulation

The most insightful evaluation of an environment can be achieved by using a detailed model of the investigated environment. Many accurate models of vehicles and other environments are freely available, and thus the effort to recreate an environment is rather low. Additionally, various simulation tools to evaluate the propagation and communication channel offer possibilities to define the materials of all components in the model. This can be a time-consuming process, but the results of the simulation are then considerably more accurate. In the following subsections, some investigated environments are described and challenges regarding their simulation are identified.

A. Car Environment

An in-car environment represents one of the harshest environments for wireless communications in the frequency range of 2.4 GHz. Lower frequencies would work better in this environment but are not an option due to the limited bandwidth. There are many sensors (e.g. vibration) that require a bandwidth with several hundred kbit/s which can be only achieved by technologies such as Bluetooth. Other technologies such as Ultra-wideband (UWB) could offer even higher data rates, but the technology yet must be optimized to be more energy-efficient, and communication on higher frequencies does not necessarily solve the physical constraints of wireless communications in harsh environments. On the contrary, some effects might be even worse. The challenge of wireless communications in this confined space is that there are different materials and components which expose wireless sensor nodes to a significant range of temperatures (-40°C - 125°C), vibrations, and dirt. Metallic surfaces block the direct path and the signal is scattered and deteriorated by many other effects described in section (II). Due to the short distances, however, there is still the possibility to achieve a reliable link with appropriate prerequisites. Different positions in the engine compartment require different transceiver structures to achieve the best possible output. Fig. 2 indicates the level of detail which is used for the simulations. All components can be given certain material properties to have a highly accurate environment.

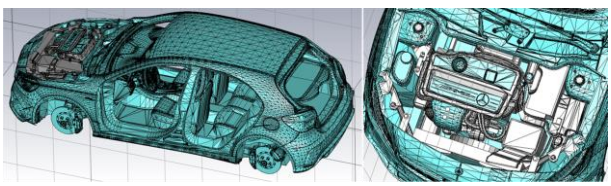


Fig. 2: Level of detail for in-vehicle simulations. Model adapted from [16].

Simulation Challenges: Simulations only consider the case of a stationary vehicle. If the vehicle is moving and the engine is running, there is a change in the whole environment. Temperature increases, the whole engine block is vibrating, and many other additional impacts must be considered. Differences in a running and stationary vehicle must be investigated, even though the first assumptions and simple tests in similar environments showed that the results will only differ slightly. External influences will be also rather low since the body of the vehicle is more or less a Faraday cage. Signals from cell towers or other mobile stations will barely interfere with internal short-range communication.

B. Traction Vehicle Environment

Another harsh environment poses traction vehicles (trains), where sensor units should be applied more frequently. There are already plenty of wired sensors incorporated by the traction vehicle manufacturer. Those sensors are mostly used to monitor states and functionality of the vehicle itself and access is only available to the manufacturer. Vehicle operators want to offer additional services for their personnel and implement active infrastructure monitoring. Such sensor systems must be mounted onto the train. For safety and security reasons there are no open interfaces to the internal sensor network of the vehicles. As a result, simple and cheap mountable wireless sensor solutions are wished for.

In the past decade, only GSM-R (Global System for Mobile Communication-Railway) was available for train communications. Just recently, LTE-R became an option, even though it has not been standardized globally for railway applications. It will just be a stopgap for a different standard called Future Railway Mobile Communication System (FRMCS), which will directly substitute GSM-R and is based on the 5G standard. However, this standard will still take several years until it is refined and can be actively used. Both, LTE and the FRMCS enable many novel sensor applications for trains.

For in-vehicle and intra-vehicle communication, many technologies have been extensively tested [17]. Most of the tested and applied sensor systems are wired, even though the commercialization of such a system will be almost impossible. The standards and safety requirements for traction vehicles are high and regulations will forbid sensor system connections to the on-board energy supply. Wireless sensor nodes, however, do not interfere with the onboard data traffic and power supply and the admission of such a system seems much more likely. In Fig. 3 the bogie of a traction vehicle is displayed.

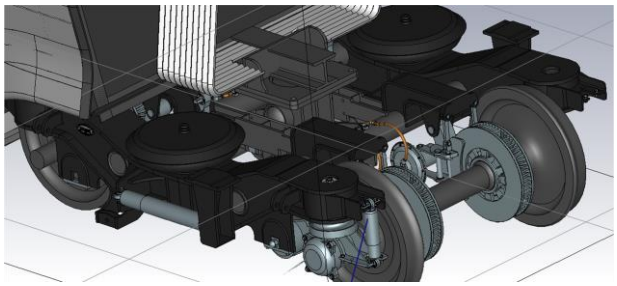


Fig. 3: Level of detail of a traction vehicle

Simulation Challenges: The challenges are quite similar to those in an in-car environment. Only stationary traction vehicles are considered and the impact of the moving vehicle, running engines, and electronics can only be estimated or simulated with immense additional effort. Generally, the radiation impact of devices on wireless communications is not that significant but can still lead to packet losses in the information transfer. In this traction vehicle scenario, many sensors are mounted on exterior components. Environmental impact by mobile stations and other areas, where interferences on the 2.4 GHz band occur, might be an issue for the wireless communication link.

C. Industrial Environment

Industrial environments are similar and often a little bit simpler than the described environments in (A) and (B). However, due to the adaptiveness of many modern factory environments, many components are movable and every change in the structure might invoke new problems. The topology cannot be fixed in this case and must be adaptive. It is possible to define classes of nodes and split them into moveable and non-movable ones. Additionally, some of the fixed transceiver structures can be substantially bigger and probably also connected to power supplies, which offers more possibilities regarding the topology. As for the other environments, accurate models can be created for industrial setups. Fig. 4 displays a smart industrial environment with several robotic assembly stations.

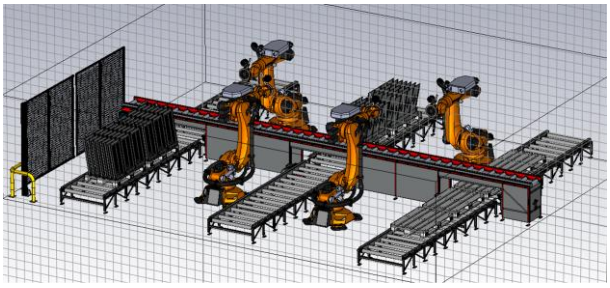


Fig. 4: Detailed industrial simulation environment. Model from [18].

Simulation Challenges: The biggest challenge is to consider all moving elements and parts of machines in this industrial context. Not only will the behavior be different for each position along a path of a mobile element, but also robotic arms might interfere with the communication if they move into other positions. These circumstances increase the complexity of an industrial environment more than initially assumed. The transceiver modules must be very robust and the antenna characteristics on those devices must withstand dynamic changes.

V. Simulating an Environment

A detailed simulation can provide indispensable information for the various communication parameters. If materials of components and interfering physical effects are modeled, the overall reliability of a communication link can be significantly improved before testing it in the real environment. This can save a lot of time and costs; thus, this approach pays off in an economical point of view. Regarding the system design, many deliberations regarding the required data rate, energy-efficiency, distance, objects, and materials in the vicinity, must be carried out. Then it is easier to identify suitable components for implementing the real system. Subsequently, we performed an exemplary demonstration/ simulation-aided evaluation for a traction vehicle. To obtain the representation for simulation, a detailed CAD model is imported into the CST Studio Suite simulation environment. Then, the wireless communication devices (antennas) are positioned and all simulation parameters are defined.

For the conducted simulations and measurements, a meandered inverted-F antenna (MIFA) was assumed since it is the most used antenna on off-the-shelf BLE devices. Initial simulations with this antenna type in the context of a traction vehicle confirm the concerns that such an antenna will not work properly for such a scenario. Antennas with diverse characteristics (more directed) have been tested and several other types (e.g. slot) could have been identified that achieve much better results in certain directions. The two types of antennas which were used for the simulations are depicted in Fig. 5.

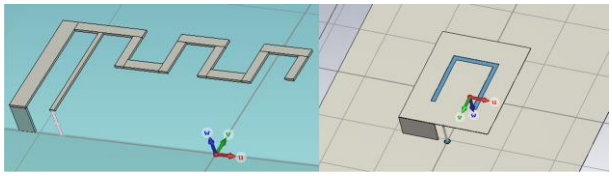


Fig. 5: Antennas used for simulations: MIFA (left), U-slot (right)

The materials for every component of the vehicle have been defined to include the influence of any metals in the simulation model. There is a very distinctive difference when material properties are used which shows the importance of using the correct measurements and properties of the various materials.

In Fig. 6, the testing environment of a traction vehicle can be seen. Two antenna positions have been defined, both at each bogie of the vehicle.

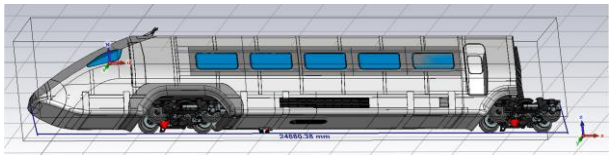


Fig. 6: Two antenna positions (red dots) on the front and rear bogie of the traction vehicle. Model adapted from [19].

Simulations with this sophisticated model yield a high attenuation for the link from the bogie to a position inside the vehicle (back door of the traction vehicle). Depending on the antenna, at least 30 dB attenuation has to be expected. The reason for this high loss is the metal shell, through which penetration of the signal is rather difficult. On average, the attenuation with MIFA antennas is even higher, in the range of 50-60 dB. The radiation pattern of MIFAs for both front and rear bogie cases are displayed in Fig. 7.

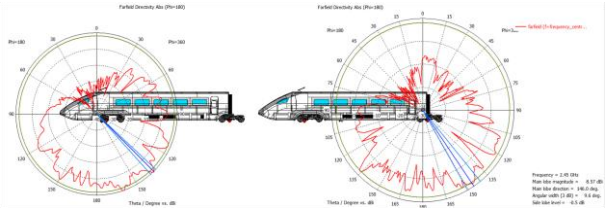


Fig. 7: Radiation pattern for common BLE antenna (MIFA) at two different locations at the front bogie and rear bogie

A wireless communication link from the bogie or some other part below the engine structure in this model is the most difficult to realize. With this specific link, it will be difficult with any type of low-energy transmitter (with according antenna) to achieve a reliable wireless link. There is only a certain amount of improvement that it is physically possible with the initial situation. Even with the mentioned slot antennas, an attenuation of 20-30 dB will occur, even if a highly directive antenna with a high gain (> 9dB) is used. Then, the required transmitting power will be too high for energy-efficient operation, and the antenna dimensions are far too large to use a compact design for the wireless sensor node. Table 1 gives an overview of the path loss, the used transmitting power, and the distance of the link.

TABLE I. COMPARISON OF ANTENNA TYPES FOR DIRECT LINK

Parameters	**Antenna Type + Position**			
	***MIFA FB*[a]**	***MIFA RB*[b]**	***U-Slot FB*[a]**	***U-Slot RB*[b]**
Path loss	-92 dB	-71 dB	-81 dB	-59 dB
Distance	16 m	3 m	16 m	3 m
TX Power	0 dBm	0 dBm	12 dBm	12 dBm

[a.] Front Bogie

[b.] Rear Bogie

In such a situation, where the highest path loss occurs due to the metal structure of the vehicle, it is worthwhile to investigate all other possibilities to optimize the communication link. For this specific case of a traction vehicle, there seems to be optimization potential concerning the topology, which has been discussed in section (III). In a traction vehicle, the sensor node positions will be fixed, thus an optimal topology can be determined in advance. To optimize the reliability of the link from the front bogie to a central transceiver node in the rear, additional nodes should be used. Considering the various materials of the train, nodes should be positioned close to spots, where penetration through the shell structure is easier than through metal. Any gaps in the hull or windows, where attenuation is significantly lower, shall be exploited in the best possible way.

Figure 8 indicates two example topologies, based on the evaluation of a parametric sweep for several locations. Those positions were picked based on the position, where the best link quality between two points could be achieved. This is a significant optimization of the overall link quality and major attenuation spots can be mostly avoided. In this simulation run, the unchanged MIFAs have been used to show the impact of positioning and intermediate nodes.

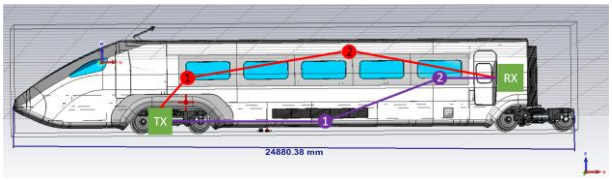


Fig. 8: Two optimized topologies (red and violet path, with the points indicating nodes) for BLE to cover both distance and reduce the path loss to a minimum.

The first node of the red path penetrates the shell at the front window. The second node is positioned inside in an unobstructed position with the LOS path to the receiver in the back. The first node of the violet path is positioned on the exterior of the train in the middle while the second node penetrates the shell at the rear window of the traction vehicle to reach the receiving device. In the simulation, the attenuation at the window is between 5-8 dB, while for the metal shell, it can be easily 30 dB or more. In comparison to the initial simulations in table 1, the attenuation of the link from the transmitter to the receiver in the rear of the traction vehicle is reduced for almost 30 dB, just by adapting the topology. The adapted results are displayed in table 2.

TABLE II. PATH LOSS OF TWO DETERMINED TOPOLOGIES

Parameters	Antenna Type + Position	
	MIFA red path	***MIFA violet path***
Path loss	-66 dB	-61 dB
Distance	4m + 8m + 7m (19 m)	9m + 6m + 4m (19 m)
TX Power	0 dBm	0 dBm

When comparing the simulated results to actual measurements [20], [21], the simulation results yield accurate results. The interior of the train also adds a little bit more attenuation to the red path than the solution (violet path) which is guided along the exterior of the train. However, the penetration on the rear windows adds more attenuation, so that the difference between both solutions is not that significant. Another benefit of adapting the topology concerns energy consumption because the required transmitting power to cover shorter distances and penetrate weaker materials is clearly lower. Consequently, the battery life of nodes is increased with this approach. The antenna has not been considered for further simulation runs, even though there is also a lot of optimization potential which could be up to 10 dB.

VI. CONCLUSION & OUTLOOK

This article takes up the cudgels for using simulations, from early phases, when designing communication systems for harsh communication environments. A few environments were discussed and analyzed. The most important communication parameters to achieve optimization in the topology using a simulation of detailed harsh environments were identified as: 1) antenna (optimal type varies for different positions in an environment), 2) topology (if the direct communication path is heavily attenuated or blocked, a different topology might be better), and 3) protocol (dependent on requirements to data rate and battery life). The application of simulations for communication system design was demonstrated using a traction vehicle scenario. The first results indicate that such an approach can indeed be used to drive the design of the communication system when detailed models are used. By considering details such as physical effects, materials, and antenna type the overall accuracy of the simulation will be close to the results in the real-world counterpart, which allows to minimizing time and costs of the design. Additionally, it allows investigating to what extend established solutions are transferable to other communication scenarios and optimize them. Future work concerns a comparison among derived solutions, as well as the evaluation of the derived systems in their real target environments, to further increase the accuracy of the simulations models.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Austrian Research Promotion Agency (FFG) (#6112792)

REFERENCES

[1] Othman, Fauzi & Shazali, Khairunnisa. (2012). Wireless Sensor Network Applications: A Study in Environment Monitoring System. Journal of Procedia Engineering. 41. 1204 – 1210. 10.1016/j.proeng.2012.07.302.

[2] M. Wollschlaeger, T. Sauter and J. Jasperneite, "The Future of Industrial Communication: Automation Networks in the Era of the Internet of Things and Industry 4.0," in *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17-27, March 2017.

[3] A. Berger, M. Pichler, D. Ciccarello, P. Priller and A. Springer, "Characterization and adaptive selection of radio channels for reliable and energy-efficient WSN," *2016 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Doha, 2016, pp. 443-448.

[4] H. Bernhard, A. Springer, A. Berger and P. Priller, "Life cycle of wireless sensor nodes in industrial environments," *2017 IEEE 13th International Workshop on Factory Communication Systems (WFCS)*, Trondheim, 2017, pp. 1-9.

[5] Min Chen, Taekyoung Kwon, Shiwen Mao, Yong Yuan, and Victor C. M. Leung. 2008. Reliable and energy-efficient routing protocol in dense wireless sensor networks. *Int. J. Sen. Netw.* 4, 1/2 (July 2008), 104–117. DOI :https://doi.org/10.1504/IJSNET.2008.019256

[6] K. Woyach, D. Puccinelli and M. Haenggi, "Sensorless Sensing in Wireless Networks: Implementation and Measurements," *2006 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Boston, MA, USA, 2006, pp. 1-8.

[7] Sana Salous, "Radio Wave Transmission," in *Radio Propagation Measurement and Channel Modelling* , Wiley, 2013, pp.35-83

[8] K. Mikhaylov, J. Tervonen, J. Heikkilä and J. Känsäkoski, "Wireless Sensor Networks in industrial environment: Real-life evaluation results," *2012 2nd Baltic Congress on Future Internet Communications*, Vilnius, 2012, pp. 1-7.

[9] Remley, K.A., Koepke, G., Holloway, C., Camell, D. and Grosvenor, C. (2009), "Measurements in harsh RF propagation environments to support performance evaluation of wireless sensor networks", *Sensor Review*, Vol. 29 No. 3, pp. 211-222. https://doi.org/10.1108/02602280910967620

[10] A. Guidara, G. Fersi, F. Derbel, M. B. Jemaa, "Impacts of Temperature and Humidity variations on RSSI in indoor Wireless Sensor Networks," Procedia Computer Science, Volume 126, 2018, pp. 1072-1081, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.08.044.

[11] Mohammad Afaneh, *Bluetooth 5 speed: How to achieve maximum throughput for your BLE application*, NovelBits, September 6, 2017. Accessed on: April 30, 2020. [Online]. Available: https://www.novelbits.io/bluetooth-5-speed-maximum-throughput/

[12] X. Li, J. Peng, J. Niu, F. Wu, J. Liao and K. R. Choo, "A Robust and Energy Efficient Authentication Protocol for Industrial Internet of Things," in *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1606-1615, June 2018.

[13] F. Bonavolontà, A. Tedesco, R. S. L. Moriello and A. Tufano, "Enabling wireless technologies for industry 4.0: State of the art," *2017 IEEE International Workshop on Measurement and Networking (M&N)*, Naples, 2017, pp. 1-5.

[14] M. Iglesias-Urkia, A. Orive, M. Barcelo, A. Moran, J. Bilbao and A. Urbieta, "Towards a lightweight protocol for Industry 4.0: An implementation based benchmark," *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, Donostia-San Sebastian, 2017, pp. 1-6.

[15] D. Kraus, K. Diwold, P. Priller, and E. Leitgeb. 2019. Achieving Robust and Reliable Wireless Communications in Hostile In-Car Environments. In 9th International Conference on the Internet of Things (IoT 2019), October 22–25, 2019, Bilbao, Spain. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3365871.3365904

[16] M. Tulio Miranda Araujo, A45 Carbon Edtion 2017, grabcad.com, 2017.

[17] Fraga-Lamas, P.; Fernández-Caramés, T.M.; Castedo, L. Towards the Internet of Smart Trains: A Review on Industrial IoT-Connected Railways. *Sensors* **2017**, *17*, 1457. https://doi.org/10.3390/s17061457

[18] J. Hudák, Robotics Workplace, grabcad.com, 2020.

[19] A. Hoti, Industrial Design (InterCity Train), grabcad.com, 2018.

[20] S. Aerts, D. Plets, L. Verloock, E. Tanghe, W. Joseph and L. Martens, "Empirical path-loss model in train car," *2013 7th European Conference on Antennas and Propagation (EuCAP)*, Gothenburg, 2013, pp. 3777-3780.

[21] M. Lerch, P. Svoboda, S. Ojak, M. Rupp and C. Mecklenbraeuker, "Distributed Measurements of the Penetration Loss of Railroad Cars," *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, ON, 2017, pp. 1-5

Distributed Ledger Technologies for IoT and Business DApps

Dejan Dolenc
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
dejan.dolenc@ltfe.org

Jan Turk
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
jan.turk@ltfe.org

Matevž Pustišek
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia
matevz.pustisek@fe.uni-lj.si

***Abstract*— The scope of the existing distributed ledger technologies is divergent in terms of technological features, as well as in their acceptance among the user and developer communities. In this paper, we propose a set of diverse criteria for the evaluation of a distributed ledger ecosystem, which does not focus only on the technological aspects of viable solutions. They also consider decentralized application development perspective and sustainability of its use. We provide a brief overview of some of the alternative distributed ledger ecosystems for the development and deployment of business and IoT decentralized applications. The set of investigated technologies and networks includes Ethereum, Hyperledger, Hedera Hashgraph, EOS, Corda, IOTA, and Multichain.**

Keywords—distributed ledger technology, blockchain, decentralized application, comparison

I. INTRODUCTION

The scope of the existing distributed ledger (DL) systems is divergent in terms of technological features, as well as in their acceptance among the user and developer communities. In the decade since their first introduction, DL research and development focused on enabling technologies and first distributed ledger technology (DLT) based decentralized applications. With the first examples of blockchain-based IoT solution deployments (e.g. in Ethereum), certain inefficiencies in these blockchain (BC) designs started appearing [1]. Micropayments for example have become almost unrealistic in many public blockchain networks, like the Bitcoin (or the Ethereum) network due to high transaction fees and long transaction confirmation times. The scalability and performance needed for the Internet of Things (IoT) (expected billions of devices) are often limited due to the size of the blockchain, limited transaction rates, and excessive latency. The most popular BC protocols endeavor to face some of these inefficiencies with functional extensions, such as state channels, sharding, and oracles. In parallel, new DL protocols are being developed, with some of the IoT requirements built-in from scratch. Both developments—IoT and DLT—are naturally seeking to be combined in common solutions, which thus provide an immense space for application development and use. However, the right approach and the selection of the appropriate technologies are far from being straightforward. Besides the differences in technological features, issues like the scalability, transaction costs, and deployment in constrained IoT devices must be considered. The same holds for the acceptance of particular blockchain technology among user and developer communities. The selection can crucially depend on the details of an intended use case, too [2].

The objective of our research was as follows:

- We propose a set of diverse criteria for the evaluation of a BC ecosystem, which do not focus only on the technological aspects of viable solutions, but also consider the Distributed Application (DApp) development perspective and sustainability of its use;
- We provide a brief overview of some of the alternative DL ecosystems for the development and deployment of business and IoT decentralized applications;

The objective of our research was not to present an introduction to DLT or to be a comprehensive survey of DLT for IoT and business. We merely took into account some of the currently most prominent technologies that kept appearing in our discussions as more or less possible alternatives to Ethereum.

The remainder of the paper is structured as follows. In Section II we present the criteria for the comparison. In Section III we provide an overview of key characteristics for the selected BC ecosystems and in Section IV summarize the key findings.

II. COMPARISON OF KEY FEATURES

A DApp [3] implementation framework consists of three key building elements, which we dubbed the DApp triplet. These elements are (i) a trusted decentralized ledger (i.e. network), (ii) trusted decentralized execution of program logic, and (iii) decentralized applications. An example of a DApp triplet is a BC network, smart contracts (SC) implemented in this network, and BC enabled front end applications, which provide user interfaces and run embedded IoT devices.

The underlying technologies and protocols strongly determine some of the features of the corresponding DL networks. The technological aspects (e.g. ledger type, consensus mechanisms) define the key performance and security limitations, as well as the environment for SCs, and software architecture for DApp development. But the same or slightly modified DL technology can be implemented with different settings and governance in different DL networks. This can result in a profoundly diverse network incarnations. The implemented network based on the technology can be e.g. public or private, which directly affects decentralization and thus the security and trust. With different ledger settings two technologically related networks may strongly differ in transaction throughput or latency. In Ethereum, a private network can provide an order of magnitude higher throughput and latency. However, the sizes of the networks and related trust might be almost incomparable.

The DLT defines a development and execution framework for smart contracts. A SC is programming code, deployed and executed in the BC network, and therefore has no central point of failure [4], [5]. For the developers it is of key importance which tools, method, and Application Programming Interfaces (API) are available for SC development and maintenance, as well as for the development of the external IoT and end-user applications, employing the DL network.

In terms of solution sustainability (developer support, software updates, development roadmap), both the organization supporting a particular technology and the community around it are very important. There are great examples of large open-source, community-based ecosystems. There are also examples of reliable technologies with future-oriented ecosystems being provided and supported by both new as well as old and renowned software companies. On the other hand, there are technology ecosystems, some of them include popular public networks and cryptocurrencies, where the solution backup and support is extremely limited

TABLE I. OVERVIEW OF KEY FEATURES

	Ethereum	**Hyperledger**	**Hashgraph**	**EOS**	**Corda**	**IOTA**	**Multichain**
Support							
By	The Ethereum Foundation	The Linux Foundation	Hedera Governing Council	Block.one	R3	IOTA Foundation	Coin Sciences Ltd
Type	Open community	Currated open community	Limited public input	Limited public input	Limited public input	Limited public input	Limited public input
Major update date	December 2019	January 2020	September 2019	October 2019	March 2020	(not clear)	April 2020
Technology							
Ledger type	chain	chain	DAG	chain	Databases of facts shared among nodes	DAG (tangle)	chain
Consensus	PoW, PoS, PoA	PoW, PoS, PoA, PoET, PBFT	ABFT	DPoS	Provisioned	PoW + Coordinator	PoW
Network types	Public, Private, Consortium	Private, Consortium	Public, Private	Public, Private	Consortium, Private	Public, Private	Private
Public network							
Public network name	Ethereum Mainnet	-	Hedera	Eos Mainnet	Corda Network	Iota Mainnet	-
Public network consensus	PoW/PoS hybrid	-	ABFT	DPoS	Provisioned	PoW + Coordinator	-
Governance	Public	-	Centralized	Public	Consortium	Centralized	-
Public network decentralization and trust	5000-7000 nodes, very high	-	10 fixed mining nodes, provided by Hedera	21 elected block producers	(not clear)	Unknown	-
Performance	Block time: 13s, Tx/s: 11, Latency: Roughly 10s	-	Latency: 3s, Tx/s: 13 for SC calls, 10000 for crypto transfers	Latency: Roughly 1s, Tx/s: unreliable, see section 3.D	100-500 seconds per 100 transactions	Latency: 2.5-3.5min, Tx/s, see section 3.F	-
Cryptocurrency	ETH	-	HBAE	EOS	-	MIOTA	-
Private networks							
Public network consensus	PoW, PoS, PoA	PoW, PoS, PoA, PoET, PBFT	ABFT	DPoS	Provisioned	PoW + Coordinator	PoW
Performance	Depends on level of centralization and protocol settings. Can be much higher than public chains.						
DApps							
SC Languages	Solidity, Vyper	Depends on project, see section 3.B	Solidity	WebAssembly	Java	-	-
Application Languages	JS, Python, Java, Go, ...		JS, Java, Go, ...	JS, Java, Swift	Java, Kotlin	JS, Go, Python, ...	-
APIs	Standard JSON RPC		JSON REST	JSON REST	Proprietary Java RPC	Proprietary JSON RPC over HTTP	Standard JSON RPC
Development support	Very good	Good	Good	Limited	Good	Poor	Poor

and might be a huge risk for real-world real business DApp development and deployment.

And last, good indicators of a DL ecosystem are its prominent use cases.

III. Distributed Ledger Ecosystems

An overview of key technological, organizational, and implementation related aspects of selected distributed ledger systems is given in Table 1. More details for the particular technologies are given in the continuation of this section.

A. Ethereum

Ethereum [6] is an open-source chain-based distributed ledger technology maintained by the Ethereum Foundation. It gets updated roughly once per year, with the last major update being Istanbul, in December 2019. To reach consensus on the network Ethereum uses the Proof of Work (PoW) consensus mechanism, switching to Proof of Stake (PoS) is already foreseen, and Proof of Authority (PoA) is available, too.

PoW is a compute-heavy confirmation algorithm that uses substantial amounts of energy to produce new blocks. PoA is a lightweight block producing algorithm running in a permissioned network, where the network administrator knows the identity of all nodes and each transaction can be authorized by a block creating node; there is no need for contention over confirmations by confirmation nodes to receive the reward.

In Ethereum, blocks are generated periodically and are limited in size by the sum of Gas of the included transactions. The amount of Gas used per transaction depends on the work required to process it, making it higher when sending or processing additional data. The minimum Gas usage per transaction is 21000.

The Ethereum blockchain technology is deployed in a large and highly decentralized public network, called mainnet. Anyone can add nodes, including mining nodes, to this network. Currently, there are about 5000-7000 nodes in the mainnet [7], [8]. The number of miners is hard to estimate, as the miners are mostly organized into mining pools [9], [10]. The currency of this network is called Ether (ETH) and it has real monetary value.

The mainnet has an average block time of 13 seconds and each block can use up to 10 million Gas. With an average Gas usage of 72000 per transaction, the mainnet is capable of 138 real-world transactions per block or 10.7 per second. The network imposes rather high latency around 10 seconds due to the high block time. For security reasons, it is advised to wait for at least 3-6 additional blocks in the chain to be sure of the transaction immutability. This imposes a delay in the order of 1 minute.

A reliable network monitor is available for the Ethereum mainnet [11].

There is also an Ethereum test network, called Ropsten. It runs similar BC protocols as the mainnet, but is smaller in scale and usually adopts the protocol upgrades first. Ether in the testnet has the same properties as in the mainnet, but no real value. Test Ether can easily be obtained through mining or test Ether faucets.

Ethereum has a good DApp development ecosystem. Many tools exist for Ethereum application development, including SC development. For this, frameworks (Truffle), Integrated Development Environments (IDE) (Remix), code validators, and network emulators are available. Smart contracts are generally written in Solidity, but Vyper and Flint are also supported. A drawback, if compared to e.g. Web development, is that reliable Solidity libraries are still not abundant.

SCs are deployed using special transactions, and function calls are simply transactions, which include coded data, and are sent to the SC address. To process contract calls that do not change the state of the smart contract, no transaction has to be sent, making them much faster and free of charge.

To access the blockchain, the apps have to connect to an Ethereum node via JavaScript Object Notation (JSON) Remote Procedure Call (RPC) over HyperText Transfer Protocol (HTTP), Inter Process Communication (IPC), or WebSockets. Both full and light Ethereum nodes are available. Well documented and supported client libraries exist in many different programming languages.

Despite a variety of existing and emerging BC technologies, Ethereum is by far the most popular platform for IoT BC applications. Among eleven cases from different application domains presented in [12], seven are based on Ethereum, and the remaining four are multiplatform (i.e., including Ethereum). The application domains include finance, games, asset management, IoT, and others. A prominent use case of Ethereum is CryptoKitties, a blockchain game on Ethereum developed by Axiom Zen that allows players to purchase, collect, breed, and sell virtual cats. Shortly after its release, the game was so popular that it caused congestion in the mainnet. Another example represents a case of an IoT device, controlled over the Ethereum network. Swether [13] is an IoT electric meter and switch, which is a part of a broader decentralized application. The DApp consists also of a corresponding smart contract deployed in the Ethereum network and a web—based Ethereum enabled user interfaces to initiate control activities.

B. Hyperledger

Rather than being a single ledger technology, Hyperledger [14] represents a suite of stable open-source frameworks, tools, and libraries for enterprise-grade blockchain deployments. It involves more than 250 teams and companies and is hosted by the Linux Foundation [15].

The Hyperledger project is comprised of various ledgers. They are all developed to support private business networks. While there are multiple projects, most use a classic ledger of chained blocks and allow for different types of nodes, therefore ensuring, that not all of the nodes need to have full chain data to communicate in the network.

Multiple projects, like Besu [16] and Sawtooth [17], support multiple confirmation algorithms, like PoW, PoA, Practical Byzantine Fault Tolerance (PBFT), or Proof of Elapsed Time PoET. PBFT [18] uses multiple confirmation nodes working together to confirm every transaction, working asynchronously. Proof of Elapsed Time uses random waiting times for confirmation nodes to select the current confirming node. Since all of the nodes are in a private non-monetized network, this is an ordering mechanism for who is next to confirm the transaction.

The software update cycle in Hyperledger varies from project to project. Since the networks are mostly maintained by companies, the updates can be more regular, than with other, public-facing blockchains. The last major milestone was achieved by the Hyperledger Fabric [19], when it reached version 2.0 on January 30th, 2020, and became the first Hyperledger project to do so.

Since Hyperledger projects are meant to support operations of various organizations, they are private or consortium-based. While there might be instances of multiple organizations accessing the same ledger or even multiple ledgers having part of the data exposed to the public, these networks are still considered private.

Network sizes vary from organization to organization and control of the network is usually centralized. The only instances of decentralization are when multiple organizations have access to the same network.

Since the Hyperledger project aims to use the advantages of blockchain technology in private and consortium networks, there is no need to incentivize participation in a network with monetary aspects. Therefore, there is no native cryptocurrency present on Hyperledger networks.

Most of the projects have network monitors provided for the network administrators to have an overview of the network [20]. There are multiple ways to test the networks. There are no public networks, but each of the projects can be deployed in a private testing environment and there are public test environments like the one in IBM cloud service.

All of the Hyperledger projects are very performant because there is only one governing body and the confirmation mechanisms can be tweaked to provide the necessary throughput. Throughput depends on the confirmation mechanic settings and can be influenced by the level of centralization (number of confirmation nodes). Latency can be less than a second since we are talking about private, possibly high bandwidth environments, with very compute capable confirmation nodes.

Every aspect of the network is under the control of the organization that is using it and therefore the trust in the network can be as high as the trust in the organization.

Hyperledger supports various SC programming languages and engines. The most prominent and widely used programming languages are Solidity and Go. There are full, light, authoritative, and other nodes. Every node is controlled and in the domain of the organization administrator. Hyperledger adapted multiple ways for API access to the nodes. Some resemble classic Web API access, other the Ethereum node access through JSON RPC over WebSocket or HTTPS, and some apply Docker. Quite extensive open source documentation is available for all projects. Smart contracts in Hyperledger can be considered immutable to the point of administrator resetting the network or deleting a certain number of blocks.

While there are many Hyperledger projects, they do not compete amongst themselves, but rather complement one another. Let's take an example of an Ethereum development collective trying to move their solution to the Hyperledger suite. They could use Hyperledger Sawtooth [17] with built-in Hyperledger Burrow [21] to run their Solidity smart contracts in the Ethereum Virtual Machine (EVM). For interaction with the Ethereum network and its monetary aspect, they could use Hyperledger Besu [16]. For the oversight over their network, they could use the Hyperledger Explorer [20]. In this manner they would use multiple Hyperledger solutions, to support their use case.

Hyperledger is hosted by the Linux foundation. Each project has its own team of collaborating organizations that vary from world-renowned IT companies to small teams, dedicated to the Hyperledger project development.

Prominent Hyperledger use cases [22] include logistic projects, financial projects, humanitarian and philanthropic projects. One of the first use cases was a seafood tracking service built on Hyperledger Sawtooth [23]. The project would track the journey of the seafood from the moment it was caught, to the moment the seafood was sold to the customer in the store. The Path of the seafood could be checked and validated in a Web interface, where the user could see every party involved in delivering the food from the sea to the store.

C. Hedera Hashgraph

Hashgraph [24] is a patented distributed ledger technology maintained by the Hedera Governing Council.

The ledger type is DAG (Directed Acyclic Graph), a tree with multiple branches. In Hashgraph, every container of transactions is incorporated into the ledger — none are discarded — so it is more efficient than blockchains. All the branches continue to exist forever and are woven together into a single whole [24]. Unlike traditional PoW blockchains, Hashgraph utilizes a form of virtual voting. It selects a single miner to choose the next block, the community of nodes running Hashgraph come to an agreement on which transactions to add to the ledger as a collective. Through gossip-about-gossip and virtual voting, the Hashgraph network comes to consensus on both the validity and the consensus timestamp of every transaction. The Hashgraph consensus algorithm has been validated as Asynchronous Byzantine Fault Tolerant (ABFT).

There is a roadmap, which started in Q3 2019, for the public Hedera Hashgraph network, with milestones scheduled 4 times per year [25]. The last major change was the Hedera mainnet (beta) made openly accessible on September 16th, 2019.

The Hedera public network [26] is built on the Hashgraph distributed consensus algorithm. It allows for creating or exchanging value, proving identity, or verifying and authenticating important data in a way that is fast, fair, energy-efficient, and secure — these advantages are almost entirely due to the underlying Hashgraph consensus algorithm [24]. There are 10 mainnet network nodes. Mirror nodes are used to access Hedera services. One cannot run a mirror node in Hedera, but there are Hedera and community mirror nodes available, with APIs for Hedera services [27]. Basic network metrics are provided on the Hedera homepage [28]. There is public cryptocurrency HBAE available. The cost of a cryptocurrency transaction can remain stable and low, currently priced at $0.0001, allowing micro-transactions (<$0.01) to be economically and technologically practical on Hedera.

There is also a public testnet [29] which provides developers with access to a free testing environment for Hedera network services. Testnets simulate the same

development environment as you would expect for mainnet. This includes transaction fees, throttles, available services, etc.

The Hashgraph consensus algorithm provides near-perfect efficiency in bandwidth usage and consequently can process hundreds of thousands of transactions per second (throttled to ten thousand Tx/s in beta) in a single shard (a fully connected, peer-to-peer mesh of nodes in a network). The performance in terms of transactions per second varies depending on the type of transaction [26]. It ranges from 13 Tx/s for smart contract transactions, over 100-1000 for consensus transactions, and up to 10000 for crypto transfer transactions. Latency is currently about 3 seconds to finality.

The Hedera network is managed by the Hedera Council and is thus not really decentralized [30].

Application developers do not run their own ledger nodes. Instead, they must rely on the API-s of the mirror nodes. The Hedera API provides access to basic ledger features, cryptocurrency accounts, consensus service, file service, and smart contracts. A Hedera smart contract is not immutable as e.g. in Ethereum. It can be changed if several parties designed by a smart contract developer agree. When deploying a smart contract on Hedera, developers can choose the contract's subsequent mutability or deploy a contract with a list of the public key of arbitrators, who can edit the code of the contract, add features, reverse particular transactions and fix bugs [31]. Smart contracts are written in Solidity [32]. The Hedera smart contract API enables you to deploy and run Solidity smart contracts on the Hedera public network. There are Hedera and community-supported Software Development Kits (SDK) for various programming languages for off-chain application development.

The Hedera Governing Council [33] consists of up to 39 term-limited and highly diversified organizations and enterprises, reflecting up to 11 unique industries, academia, and non-profits globally. Council members are committed to governing software changes while bringing stability and continued decentralization to the public network.

Prominent Hedera use cases include payments, tokenized assets, and managing credentials [34]. Power Transition is a Microgrid Management Platform to transform how energy is used, managed, and traded by governments, businesses, and individuals. Acoer provides open-source solutions for pharma and life sciences to improve the transparency and efficiency of communication in healthcare.

D. EOS

EOS [35] is a blockchain-based distributed ledger technology, made by Block.one [36]. New features and fixes are being developed for EOS quite quickly. There is a new release every 2-6 months. The last one, version 2.0.0, released in January 2020, was a major release, which introduced faster smart contracts, consensus protocol updates, and more.

While transactions themselves are free in EOS, you still CPU and memory are not, but you eventually get those tokens back. Regular transactions are cheaper and contract calls are more expensive, depending on what they do. Miners and the network are paid for using the inherent inflation of the network. EOS uses a fairly traditional system of blocks, in which new blocks are constantly being generated. The exact block rate depends on network settings. To reach consensus in the network, EOS uses Delegated Proof of Stake (DPoS). In this method, there are 21 block producers, which take turns producing blocks in a round-robin fashion. To select these 21 block producers, users vote on a list of candidates. The more EOS currency a user has, the more their vote is worth. Users can also allow others to vote on their behalf [37]. Anyone can start their own EOS node and, if voted for by the community, can become a block producer. EOS also offers a public testnet [38].

Currently, the EOS Mainnet is running at 2 blocks per second, giving it very low latency, and is advertised to have extremely high throughput, reportedly having reached a maximum of almost 4000 transactions per second. It is important to note that EOS counts many operations as separate transactions, so a single transaction in other blockchain networks, for example, Ethereum, would count as multiple transactions in EOS, even though the same operations were performed.

EOS trades some of the ultra-decentralized nature of other blockchains for transaction throughput. Instead of the miner of the next block being unpredictable, the miner is known because it has the status of a block producer, but it can lose this status very easily. Some consider this method to be too centralized and insecure [39]. The only officially supported EOS nodes are full nodes, but the protocol does allow for light nodes and there are unofficial implementations available [40]. Clients can communicate with these nodes using a REST API [41].

Officially supported client libraries exist for JavaScript, Java, and Swift [42]. Only auto-generated documentation exists [43] and the community is there, but it's small and not very active. Smart contracts in EOS run WebAssembly, meaning that a wide variety of languages is supported [44]. There are also many development tools available [45]. All EOS-related code is open-sourced and developed by the community, but Block.one [36] is still the company behind it.

Many DApps exist that use EOS, most of them gambling oriented. A list of the most popular ones is available [46].

E. Corda

Corda [47] is open-source distributed ledger technology developed and maintained by the R3, a technology company founded in 2014. R3 gained prominence in 2015 when a consortium of banks joined the initiative. Corda was first released in 2016 and there we 4 major releases since [48]. It was updated to the current version 4.4 in March 2020 [49].

In Corda, there is no single central store of data. Instead, each node maintains its own database of those facts that it is aware of. The facts that a node knows about are those that it is involved with. The network uses point-to-point messages, which are sent on a need to know basis (lazy propagation). Determining whether a proposed transaction is a valid ledger update involves reaching two types of consensus. Validity consensus is checked by each required signer before they sign the transaction. Uniqueness consensus is only checked by a notary service [50].

A Corda network is a publicly-available peer-to-peer network of nodes. Each node represents a legal entity, runs Corda software, and is operated by network participants. It is identified by a certificate and will also be identifiable on a network map. The networks are semi-private. To join a

network, a node must obtain a certificate from the network operator. This certificate maps a well-known node identity to a real-world legal identity and the corresponding public key. There is no public cryptocurrency in (public) Corda networks [50]. Corda Testnet [29] is an open public network of Corda nodes on the internet. It is designed to be a complement to the Corda Network. It is designed for "non-production" use, including but not limited to CorDapp development, multi-party testing, demonstration and showcasing of applications and services, learning, training, and development. The Corda Testnet is based on the same technology as the main Corda Network, but can be joined on a self-service basis through the automated provisioning system.

For several reasons, Corda's performance is hard to be defined in a way comparable to the number of transactions per second known in e.g. Ethereum. First, Corda supports different types of transaction flows, which differ in the inclusion in the ledger. One transaction can also hold multiple payments. Second, the throughput changes with the number of output states per transaction and how that allows the node to achieve a greater number of Corda states to be transacted per second. Third, it runs as a private network and the performance strongly depends on different host configurations, too [51]. Some documented representative cases of Corda deployments indicate up to several 1000 transactions per second [52]. In a real use case, where batches of 100 transactions with different size of the attached data, the measured latency was around 100-500 seconds per 100 transactions [53].

The Corda networks are private or consortium-based, therefore the decentralization depends on a particular case.

After starting quickly on Corda one can easily migrate to Corda Enterprise, as business requirements evolve. A network node is a Java Virtual Machine (JVM) run-time with a unique network identity running the Corda software. The node has two interfaces with the outside world, a network layer, for interacting with other nodes, and the RPC, for interacting with the node's owner. The node's functionality is extended by installing CorDapps in the plugin registry [50].

Rich, up-to-date, and well-structured documentation, including code examples, is available to the developers. The Corda API is well documented and supports every aspect of Dapp development [54]. There is a rich set of development and monitoring tools available [55].

SCs in Corda are agreements whose execution is both automatable by computer code working with human input and control, and whose rights and obligations, as expressed in legal prose, are legally enforceable. The smart contract links business logic and business data to associated legal prose to ensure that the financial agreements on the platform are rooted firmly in law and can be enforced in the event of ambiguity, uncertainty or dispute. The code in Corda is written using Kotlin, a programming language from JetBrains that targets the JVM and JavaScript. The virtual machine selected for contract execution and validation is an augmented and radically more restrictive version of the Java Virtual Machine, which enforces not only security requirements but also deterministic execution [56].

To govern development and networks, a separate entity called Corda Network Foundation has been set up, using a not-for-profit legal entity type known as a Stichting. This type is suited for governance activities, able to act commercially, with limited liability but no shareholders, capital, or dividends. Its constitution is defined in a set of Articles of Association and By-laws. A Governing Board of 11 representatives has been already established. A Technical Advisory Committee is comprised of representatives of Participant organizations. A Governance Advisory Committee is comprised of representatives of Participant organizations. The Network Operator charges the Foundation a reasonable sum for providing network and administration services, paid by the Foundation. Participation is open to any legal entity participating in Corda Network and is independent of R3 alliance membership [57].

The Corda Enterprise is proven to meet the security, scalability, and support requirements of complex organizations, and is now the de facto standard in financial services. It is therefore used primarily in business [58]. The SPUNTA Banca DLT solution, built on Corda Enterprise by SIA, NTT Data, and governed by the Italian Banking Association is live with ~80% of the Italian banking industry now preparing to use a live blockchain application to transform the interbank reconciliation process. CIeNET proposed a blockchain-based decentralized solution for MNP (DMNP), which leverages the Corda blockchain technologies to streamline the number portability flow. With Corda, all the mobile numbers are managed in decentralized ledgers and all the operators are connected into the same Corda networks. The mobile numbers remain synchronized among all the mobile operators. This enables automated number portability processes, shortening and reducing service outages.

F. IOTA

IOTA [59] is an open-sourced distributed ledger technology developed by the IOTA Foundation. It uses a special kind of ledger, called Tangle [60]. It works very differently than other distributed ledger technologies.

In IOTA, the sender of a transaction has to do the PoW. Transactions are then reinforced when other transactions reference them. At the moment, transactions only become truly valid when the network Coordinator, which is a special node operated by the IOTA Foundation, confirms them [60].

There is a public main network and a public test network. It is also possible to set up private networks, but it is tricky because you need to set up a Coordinator node [61]. Note that when connecting to the IOTA network, you have to manually add neighbors. A network monitor and explorer are available. It is impossible to know exactly how many nodes there are on the IOTA main public network, but there are sites with listed nodes so that people setting up their own nodes can connect to them [62]. Because of IOTA's unique ledger style, it has some unique properties. Theoretically, the transaction throughput increases, and the latency decreases as the number of nodes and users increases. Currently, the IOTA main network should be able to handle between 500 and 800 transactions per second [63] with a latency of 2.6 minutes [64].

IOTA networks use a public cryptocurrency MIOTA. Transactions are essentially free, but the transaction sender must submit the proof of work.

Support for smart contracts is being worked on [65], but is not available yet, meaning that IOTA can only be used for

storage and transmission of data, but not processing. This makes it, for the moment, not usable for building DApps.

G. Multichain

Multichain [66] is an open-source platform for private blockchains, which promises a rich set of features including extensive configurability, rapid deployment, permissions management, native assets, and data streams [67]. It was first released in June 2015 by Coin Sciences[68]. Updates in the current 2.0 version are released about 4 times per year. The last update was available in April 2020 for version 2.0.6 Demo for Linux. The last major update in the Multichain git repository was at the end of 2018.

Multichain provides maximal compatibility with the bitcoin ecosystem, including the peer-to-peer protocol, transaction/block formats, and Bitcoin Core APIs/runtime parameters. Multichain is a platform that is entirely dedicated to the creation and deployment of private blockchains. It can be used within or between organizations [69]. There is no public cryptocurrency in (public) Multichain networks, but it supports multi-currency tokens for peer to peer transactions. There is an open-source network monitor available on GitHub. The online explorer was not responding at the time of writing of this paper. No relevant public test networks seem to be available.

Being a private network the performance is strongly related to the network and node configuration details. Some sources mention transaction rates from 100-1000 Tx/s, however, most of these sources seem to be outdated (2-3 years old) [70]. The level of decentralization depends on a particular network case. Multichain JSON-RPC API is available in the node [71].

The company behind the Multichain is Coin Sciences Ltd [68]. Multichain does not support smart contracts and is thus not appropriate for DApps. Besides, its glory days seem to have passed with the 2018 cryptocurrency crash.

IV. Conclusion

Ethereum, Hyperledger Project, and Hedera Hashgraph are in our opinion top choice ecosystems for the development of sustainable decentralized applications for business and IoT. They all provide a variety of tools, documentation, and solid developer support. Ethereum is distinguished by a renowned, highly decentralized public network, with to some extent limited performance. The performance can be vastly improved if its technology is applied in a private network. Hyperledger focuses exclusively on private DL networks. The project provides a suite of stable frameworks, tools, and libraries for enterprise-grade blockchain deployments. Hedera is a public network based on the Hashgraph DLT. The network seems to be more performant than Ethereum, with lower transaction costs. However, the governance of Hedera is highly centralized, so trust in the network needs to be evaluated with care. EOS lacks a good developer support and honest performance measurement results. There are private as well as public EOS networks. Like in Hedera the latter seems to be very centralized.

IOTA has always been a promising concept in terms of extremely performant DL network. Unfortunately, the lacking developer support and the absence of smart contracts make it inappropriate for advanced DApps. It seems that the IOTA Foundation did not make the appropriate shift from the promising (crypto) technology in 2018. The same holds for Multichain. To found your DApp solution on these two seems complicated due to technology limitations and risky due to lacking reliable and long term developer ecosystems.

And finally, Bitcoin and XRPL were not even included in our study, since they are inappropriate for modern decentralized applications. The only relevant application of Bitcoin protocol is (and will be in the foreseeable future) the Bitcoin network and the corresponding cryptocurrency. The same holds for XRLP, too.

Acknowledgment

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-024, ICT4QoL—Information and Communications Technologies for Quality of Life).

References

[1] X. Zheng, Y. Zhu, and X. Si, ‘A Survey on Challenges and Progresses in Blockchain Technologies: A Performance and Security Perspective’, *Applied Sciences*, vol. 9, no. 22, p. 4731, Jan. 2019, doi: 10.3390/app9224731.

[2] M. Pustišek, A. Umek, and A. Kos, ‘Approaching the Communication Constraints of Ethereum-Based Decentralized Applications’, *Sensors*, vol. 19, no. 11, p. 2647, Jan. 2019, doi: 10.3390/s19112647.

[3] W. Cai, Z. Wang, J. B. Ernst, Z. Hong, C. Feng, and V. C. M. Leung, ‘Decentralized Applications: The Blockchain-Empowered Software System’, *IEEE Access*, vol. 6, pp. 53019–53033, 2018, doi: 10.1109/ACCESS.2018.2870644.

[4] A. M. Antonopoulos and G. Wood, *Mastering Ethereum - Building Smart Contracts and DApps*. O’Reilly Media, 2018.

[5] C. Cachin, ‘Architecture of the Hyperledger Blockchain Fabric’, in *Workshop on Distributed Cryptocurrencies and Consensus Ledgers*, Chicago, Illinois, USA, Jul. 2016, Accessed: Apr. 28, 2020. [Online]. Available: https://www.zurich.ibm.com/dccl/#program.

[6] ‘Ethereum White paper’, Aug. 28, 2019. https://github.com/ethereum/wiki (accessed Aug. 28, 2019).

[7] ‘Clients - ethernodes.org - The Ethereum Network & Node Explorer’. https://ethernodes.org/?synced=1 (accessed May 06, 2020).

[8] ‘Ethereum Node Tracker | Etherscan’. https://etherscan.io/nodetracker# (accessed May 06, 2020).

[9] ‘Top Miners over the last 24h - etherchain.org’. https://etherchain.org/charts/topMiners (accessed May 06, 2020).

[10] ‘Home - Ethermine - Ethereum (ETH) mining pool’, *Ethermine*. https://ethermine.org/ (accessed May 06, 2020).

[11] etherscan.io, ‘Ethereum (ETH) Blockchain Explorer’, *Ethereum (ETH) Blockchain Explorer*. http://etherscan.io/ (accessed May 06, 2020).

[12] K. R. Ozyilmaz and A. Yurdakul, ‘Designing a Blockchain-Based IoT With Ethereum, Swarm, and LoRa: The Software Solution to Create High Availability With Minimal Security Risks’, *IEEE Consumer Electronics Magazine*, vol. 8, no. 2, pp. 28–34, Mar. 2019, doi: 10.1109/MCE.2018.2880806.

[13] M. Pustišek, N. Bremond, and A. Kos, ‘Electric Switch with Ethereum Blockchain Support’, *IPSI TIR*, vol. 14, no. 1, pp. 21–28, Jan. 2018.

[14] ‘Hyperledger’. https://www.hyperledger.org/ (accessed Apr. 09, 2019).

[15] ‘The Linux Foundation – Supporting Open Source Ecosystems’, *The Linux Foundation*. https://www.linuxfoundation.org/ (accessed May 09, 2020).

[16] ‘Hyperledger Besu’, *Hyperledger*. https://www.hyperledger.org/projects/besu (accessed May 10, 2020).

[17] ‘Hyperledger Sawtooth’, *Hyperledger*. https://www.hyperledger.org/projects/sawtooth (accessed May 10, 2020).

[18] M. Castro and B. Liskov, ‘Practical Byzantine fault tolerance’, in *Proceedings of the third symposium on Operating systems design and implementation*, New Orleans, Louisiana, USA, Feb. 1999, pp. 173–186, Accessed: May 09, 2020. [Online].

[19] 'Hyperledger Fabric', *Hyperledger*. https://www.hyperledger.org/projects/fabric (accessed May 10, 2020).

[20] 'Hyperledger Explorer', *Hyperledger*. https://www.hyperledger.org/projects/explorer (accessed May 09, 2020).

[21] 'Hyperledger Burrow', *Hyperledger*. https://www.hyperledger.org/projects/hyperledger-burrow (accessed May 10, 2020).

[22] 'Case Studies', *Hyperledger*. https://www.hyperledger.org/resources/case-studies (accessed May 09, 2020).

[23] 'Seafood Case Study in Supply Chain Traceability Using Blockchain Technology | Sawtooth Distributed Ledgers'. https://sawtooth.hyperledger.org/examples/seafood.html (accessed May 09, 2020).

[24] H. Hashgraph, 'What is Hashgraph consensus?', *Hedera Hashgraph*, May 06, 2020. http://www.hedera.com/learning/what-is- Hashgraph-consensus (accessed May 06, 2020).

[25] H. Hashgraph, 'Roadmap', *Hedera Hashgraph*, May 05, 2020. http://www.hedera.com/roadmap (accessed May 06, 2020).

[26] 'Mainnet'. https://docs.hedera.com/guides/mainnet (accessed May 06, 2020).

[27] 'Mirror Nodes'. https://docs.hedera.com/guides/mainnet/nodes/mirror-nodes (accessed May 06, 2020).

[28] H. Hashgraph, 'Hello future', *Hedera Hashgraph*, May 06, 2020. http://www.hedera.com (accessed May 06, 2020).

[29] 'Joining Corda Testnet', Jan. 08, 2020. https://docs.corda.net/docs/corda-enterprise/4.4/network/corda-testnet-intro.html (accessed May 06, 2020).

[30] 'Who can run a Hedera mainnet node?', *Hedera Help*. http://help.hedera.com/hc/en-us/articles/360000665338 (accessed May 06, 2020).

[31] 'Hedera Hashgraph vs Blockchain | Comparison', *Software Development Company*, Mar. 29, 2018. https://www.leewayhertz.com/Hashgraph-vs-blockchain/ (accessed May 06, 2020).

[32] H. Hashgraph, 'Smart Contract', *Hedera Hashgraph*, May 06, 2020. http://www.hedera.com/smart-contract (accessed May 06, 2020).

[33] H. Hashgraph, 'Council', *Hedera Hashgraph*, May 05, 2020. http://www.hedera.com/council (accessed May 06, 2020).

[34] 'Hedera Hashgraph users'. https://www.hedera.com/users (accessed May 06, 2020).

[35] 'EOSIO - Blockchain software architecture', *EOSIO*. https://eos.io/ (accessed May 09, 2020).

[36] 'Block.one - High Performance Blockchain Solutions', *Block.one*. https://block.one/ (accessed May 09, 2020).

[37] G. Konstantopoulos, 'Understanding Blockchain Fundamentals, Part 3: Delegated Proof of Stake', *Medium*, Feb. 06, 2020. https://medium.com/loom-network/understanding-blockchain-fundamentals-part-3-delegated-proof-of-stake-b385a6b92ef (accessed May 09, 2020).

[38] 'EOSIO Testnet'. https://testnet.eos.io/ (accessed May 09, 2020).

[39] 'DPOS Blockchains: Is Decentralization At Stake?', *Crypto Briefing*, Jul. 25, 2019. https://cryptobriefing.com/dpos-blockchains-decentralization-stake/ (accessed May 09, 2020).

[40] eosforce, *eosforce/eos-light-node*. 2020.

[41] 'EOSIO Development Documentation - Nodeos RPC API Referenc', *EOSIO Developer Portal*. https://developers.eos.io/welcome/latest/reference/nodeos-rpc-api-reference (accessed May 09, 2020).

[42] 'Build on EOSIO', *EOSIO*. https://eos.io/build-on-eosio/ (accessed May 09, 2020).

[43] 'EOSIO Development Documentation - Api Reference', *EOSIO Developer Portal*. https://developers.eos.io/manuals/eosjs/latest/API-Reference/index, https://developers.eos.io/manuals/eosjs/v20.0/API-Reference/index (accessed May 09, 2020).

[44] appcypher, *appcypher/awesome-wasm-langs*. 2020.

[45] 'Community Developer Tools'. https://developers.eos.io/welcome/latest/community-developer-tools/index, https://developers.eos.io/welcome/v2.0/community-developer-tools/index (accessed May 09, 2020).

[46] 'Top EOS Dapps', *DappRadar*. https://dappradar.com/rankings/protocol/eos (accessed May 09, 2020).

[47] 'Corda | Open Source Blockchain Platform for Business', *Corda*. https://www.corda.net/ (accessed May 06, 2020).

[48] 'History of R3 Blockchain Platform & Corda Enterprise', *R3*. https://www.r3.com/history/ (accessed May 06, 2020).

[49] 'Release notes', Jan. 08, 2020. https://docs.corda.net/docs/corda-os/4.4/release-notes.html (accessed May 06, 2020).

[50] 'Consensus', *Corda*, Jan. 08, 2020. https://docs.corda.net/docs/corda-os/4.4/key-concepts-consensus.html (accessed May 06, 2020).

[51] 'Sizing and performance', Jan. 08, 2020. https://docs.corda.net/docs/corda-enterprise/4.4/node/sizing-and-performance.html (accessed May 06, 2020).

[52] 'Throughput — A Corda story', *Corda*, Jan. 03, 2019. https://www.corda.net/blog/throughputa-corda-story/ (accessed May 06, 2020).

[53] 'Experimenting with Corda Attachments', *Corda*, Mar. 02, 2020. https://www.corda.net/blog/experimenting-with-corda-attachments/ (accessed May 06, 2020).

[54] 'API: Contracts', Jan. 08, 2020. https://docs.corda.net/docs/corda-os/4.4/api-contracts.html (accessed May 06, 2020).

[55] 'Corda Tools', Jan. 08, 2020. https://docs.corda.net/docs/corda-enterprise/4.4/tools-index.html (accessed May 06, 2020).

[56] 'Technical difference between Ethereum, Hyperledger fabric and R3 Corda'. https://medium.com/@micobo/technical-difference-between-ethereum-hyperledger-fabric-and-r3-corda-5a58d0a6e347 (accessed May 06, 2020).

[57] 'Corda Network'. https://corda.network/governance/index/ (accessed May 06, 2020).

[58] 'Enterprise Blockchain Use Cases | Blockchain for Business', *R3*. https://www.r3.com/customers/ (accessed May 06, 2020).

[59] 'The Next Generation of Distributed Ledger Technology'. https://www.iota.org/ (accessed May 06, 2020).

[60] A. Gal, 'The Tangle: an Illustrated Introduction', *Medium*, Apr. 09, 2018. https://blog.iota.org/the-tangle-an-illustrated-introduction-4d5eae6fe8d4 (accessed May 06, 2020).

[61] 'IOTA networks | Network | Getting Started | IOTA Documentation'. https://docs.iota.org/docs/getting-started/0.1/network/iota-networks (accessed May 06, 2020).

[62] 'IOTA Nodes'. https://iota-nodes.net/statistics (accessed May 06, 2020).

[63] S. #cryptocurrency • 3 Y. Ago, 'Transaction Speed - Bitcoin, Visa, Iota, Paypal', *Steemit*, Jul. 23, 2017. https://steemit.com/cryptocurrency/@steemhoops99/transaction-speed-bitcoin-visa-iota-paypal (accessed May 06, 2020).

[64] 'TangleMonitor - Live visualisation and metrics of the IOTA Tangle'. https://tanglemonitor.com/ (accessed May 06, 2020).

[65] E. Hop, 'The State of Qubic', *Medium*, Apr. 13, 2020. https://blog.iota.org/the-state-of-qubic-63ffb097da3f (accessed May 06, 2020).

[66] 'MultiChain | Open source blockchain platform'. https://www.multichain.com/ (accessed May 06, 2020).

[67] *MultiChain/multichain*. MultiChain, 2020.

[68] 'About Coin Sciences | MultiChain'. https://www.multichain.com/about-coin-sciences-ltd/ (accessed May 06, 2020).

[69] 'MultiChain vs Ethereum : The Ulitmate Face Off', *ReadWrite*, Jan. 11, 2020. https://readwrite.com/2020/01/10/multichain-vs-ethereum-the-ulitmate-face-off/ (accessed May 06, 2020).

[70] 'Tips for performance optimization | MultiChain'. https://www.multichain.com/developers/performance-optimization/ (accessed May 06, 2020).

[71] 'MultiChain JSON-RPC API commands | MultiChain'. https://www.multichain.com/developers/json-rpc-api/ (accessed May 06, 2020).

TinyI2C - A Protocol Stack for connecting Hardware Security Modules to IoT Devices

Thomas Fischer*†, Dominic Pirker*†, Christian Lesjak†, Christian Steger*
Email: {thomas.fischer3, dominic.pirker, christian.lesjak}@infineon.com, steger@tugraz.at
*Institute for Technical Informatics, Graz University of Technology, Austria
†Design Center Graz, Infineon Technologies Austria AG

***Abstract*—To enhance the security of devices in the Internet of Things, devices are augmented with Hardware Security Modules (HSMs). To connect HSMs to their hosting devices, serial interfaces, e.g. I2C, are used. On top of these interfaces, a protocol stack is utilized to establish a reliable communication channel. HSM vendors, such as Microchip, NXP, and Infineon, use protocols that differ in regard of provided features, complexity, and efficiency. These protocols are either complex to implement, or lack certain features. In the first case, this leads to significant system integration effort, in the latter, the HSM's reliability and interchangeability suffers.**

In this paper, we perform an evaluation of state-of-the-art solutions, GlobalPlatform APDU Transfer over I2C, Microchip cryptoauthlib, and the Infineon I2C Protocol Stack. Based on this evaluation, we propose TinyI2C, a lightweight communication protocol stack. It is designed to allow simple implementations, while providing equivalent core features as state-of-the-art solutions, including reliability and packet fragmentation. Major design goals were to create a symmetric protocol, where code can be shared between both peers, which is not the case in state-of-the-art solutions. In addition, we add features, such as packet streaming support, to make the protocol suitable for Remote-Procedure-Call (RPC) based frameworks. Finally, we show a proof-of-concept and evaluate the achieved performance.

***Index Terms*—I2C, Protocol Stack, Reliability, GlobalPlatform, Microchip, Infineon, IoT, HSM**

I. Introduction

Information security is a key aspect for IoT devices to be successful on the market. To proof the identity of devices to their corresponding communication peers, digital signatures are utilized for authentication. The creation of a signature requires the possession of secret key material that has to be stored on the device. In order to prevent identity theft, manufacturers augment devices with Hardware Security Modules (HSMs). Signature creation and storing key material is delegated towards the HSM. This measure significantly impedes attacks to retrieve secret key material from devices, such as cache [7] or other side-channel attacks [13].

Figure 1 depicts a typical system architecture for an IoT device. It consists of a host controller, running the main application, and a HSM, hosting security sensitive logic. In this work, we focus on the connection between host controller and HSM. To interconnect these two parts, serial interfaces, such as SPI or I2C, are used. Since transmission errors can occur on any physical link, reliability is required on the communication channel. Unlike hardware solutions for reliability on an I2C bus, as depicted in [4], we focus on pure software solutions. In addition, the limited frame size of interface module implementations has to be considered. Therefore, support for packet fragmentation is mandatory to support arbitrary host controllers with a protocol. Several established industry standards for protocol stacks, dedicated for HSMs, exist. For example, Trusted Platform Modules (TPMs), a variant of HSMs utilized in Personal Computers (PCs) and Notebooks, use a standardized protocol stack, specified by the Trusted Computing Group (TCG) [15]. For IoT devices several solutions exists. Different vendors use different protocols that differ in regard of features, complexity, and efficiency.

In this paper, we evaluate state-of-the-art solutions for protocol stacks that are utilized for HSMs in the IoT area. Based on these findings, we propose TinyI2C, a protocol stack optimized for resource-constraint IoT devices. While retaining the features of state-of-the-art solutions, reliability and packet fragmentation, the main design goal is to allow simple implementations with a low memory footprint. We show a proof of concept (PoC) and evaluate it by comparing key indicators, such as transaction speed and required memory. Main contributions of this work are:

- Evaluation of state-of-the-art I2C protocol stacks for connecting HSMs to IoT devices
- Proposal of a simple and light-weight protocol stack
- Proof-of-Concept in the C programming language
- Evaluation of the PoC in terms of performance

II. State of the Art

This section gives an overview of state-of-the-art I2C protocol stacks for IoT devices. We extracted a list of protocol

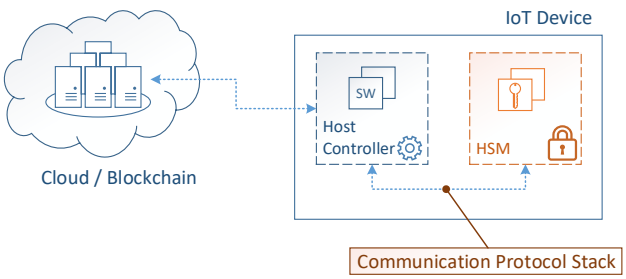


Fig. 1. Typical IoT Device System Architecture

features to conduct a structured comparison. These features appear in at least one protocol. We differentiate between features related to the physical layer, data link layer, and features that can be placed in any layer. For the physical layer the list of extracted features includes:

- Interaction type
- Communication Interface Parameter (CIP) configuration
- Timeout handling

Interaction type is defined as the way the protocol interacts with the physical interface. We observed two types, a register-based approach and a request-response state machine. The concept of registers is typical for I2C sensors, e.g. for MEMS motion sensors as depicted in [11]. The idea is that the device provides a set of registers that can be read and written via the external interface. Next, the CIP configuration defines the initializion process of physical interface parameters. The simple approach is to pre-provision static settings to both devices, the more complex is to have a parameter negotiation in the protocol. This allows auto-configuration of the host controller's interface settings, which increases interchangeability of protocol implementations. Finally, the timeout handling defines, how the host controller decides when to abort waiting for the I2C slave device.
For the data link layer the following aspects have been identified:

- Interaction type
- Reliability (checksums and sequence numbers)
- Header size (encoding and efficiency)
- Implementation symmetry

The interaction type has the same meaning as for the physical layer. A request-response state-machine is the simplest option, more complex designs support packet streaming. Reliability depends on mechanisms, such as the use of checksums, to protect the integrity of messages, and sequence numbers, to detect lost and duplicate frames. If sequence numbers are used, the receiver must acknowledge all valid messages, otherwise retransmission occurs. The concept of window size allows transmission of multiple frames until an acknowledgement has to be received. With this feature, speed is improved, and the number of required acknowledgments is reduced. To allow reliable packet streaming, the support of sequence numbers in combination with a window size of at least one is mandatory. Besides that, header size is an important aspect, since it relates to transmission overhead imposed by the protocol. Finally, implementation symmetry leads to reduced implementation and maintainance effort. Code can be shared between both peers, if a protocol design has this aspect.

Additional features, that are not always designed into the same layer are:

- Packet fragmentation
- Logical channels

Packet fragmentation is required when the maximum frame size of the physical interface is exceeded. That occurs if interface modules or drivers have limited buffers or length registers. Lastly, the support of logical channels allows multiplexing traffic between several applications running on a single device.

The extracted features are also part of other well-known protocols. For instance, the Transmission Control Protocol (TCP) [8], that is widely-used in the Internet, provides reliability by using checksums and sequence numbers. The Internet Protocol (IP) provides packet fragmentation in case the Maximum Transmission Unit (MTU) is exceeded. And the Bluetooth Low Energy (BLE) protocol [3] provides reliability functionality in the Link Layer with sequence numbers and acknowledgments. These examples are more complex then the protocols intended for HSMs, but there is also more research available, especially regarding reliability and stability, e.g. in [12], the authors analyze the reliability of TCP with statistical methods. With the collected features, we conduct a structured comparison of three industry standards that are designed to be used with HSMs.

A. Infineon I2C Protocol Stack

The Infineon I2C Protocol Stack (IFX-I2C) [14] is a proprietary communication protocol for serial interfaces. Currently only I2C is officially supported, however the design is generic and can be adopted to other serial interfaces as well. The design is based on the ISO/OSI reference model and consists of several layers as depicted in Figure 2.

1) Physical Layer: The lowest layer is the physical layer, consisting of a set of registers for I2C. One of these registers, the DATA register, is intended for relaying data frames to the upper layer. Other registers provide triggers for additional features. These include a software-reset (SWR) mechanism, reading Communication Interface Parameters (CIP) and changing the I2C address of the device. Registers are accessed using the Late-Acknowledge Algorithm. This specifies that the I2C bus master performs polling until the slave device is ready to answer.

2) Data Link Layer: The data link layer provides a reliable communication channel, by adding CRC checksums and sequence numbers to frame headers. As in TCP, there is window size support. Due to the size of the bit-field dedicated for the sequence number, the maximum window size is limited to two frames. The protocol distinguishes between data and control frames. Control frames have the same header structure as data frames, but carry no payload. Typically, control frames are only needed to acknowledge the last frame in a transmission chain, or to acknowledge a frame, in case a timeout would occur until the next data frame is available from the higher layer. To allow recovery from severe error conditions, a mechanism for re-synchronization is defined.

3) Network and Transport Layers: The transport and network layer are strongly interconnected since they share the same header fields to allow an efficient encoding. The network layer supports multiple logical channels but this feature is currently unused in Infineon's products utilizing this protocol. The transport layer supports packet fragmentation, a bit-field consisting of three bits is used to encode single, first, intermediate, and last frames. An additional value is used to

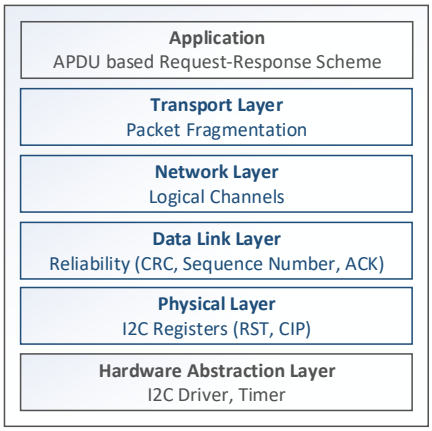


Fig. 2. Infineon I2C Protocol Stack

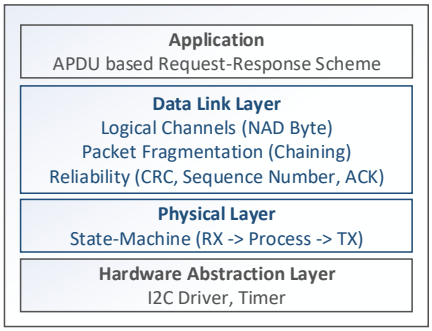


Fig. 3. GlobalPlatform APDU Transfer over I2C

indicate a chaining error. Since the data link layer provides a reliable communication link, such errors should only occur in rare cases. These include faulty protocol implementations, and when a frame is corrupt but the CRC checksum is still valid due to a CRC value collision.

The API towards the application and interface drivers is not specified in the Standard. Typically, an APDU based request-response communication scheme is used on top of this protocol.

B. GlobalPlatform APDU Transfer over I2C

The GlobalPlatform APDU Transfer over SPI / I2C specification [1] is a joint undertaking of various companies, including NXP, Infineon, and ST. It aims to standardize how Secure Elements (SEs) are connected to their hosting devices. The Standard is based upon the ISO/IEC 7816-3 [10] specification for chip cards, which specifies the T1 protocol. The GlobalPlatform specification provides a delta to the T1 protocol; therefore the new protocol is called T1'. I2C and SPI are defined as supported physical layers, in this paper we focus on I2C.

1) Physical Layer: Figure 3 shows the components of this protocol stack. The physical layer consists of a state-machine that limits the interaction to a request-response scheme. After power up, the device is in receive state and transits to processing state once a frame has been received. While in this state, the device ignores all I2C requests, as in the Late-Acknowledge Algorithm. In order to know when the processing is complete, the hosting device either performs polling, or an interrupt line is used as indication. After processing is complete, the transmit state is entered. Then, the hosting device must retrieve the answer from the SE to trigger the transition into the receive state. This behavior implies that sending new requests is only possible after receiving the last answer.

2) Data Link Layer: The data link layer provides reliability by adding a CRC checksum and sequence number to the frame headers. A single bit sequence number is used, therefore the window size is fixed to one frame. A dedicated bit in the protocol header supports packet fragmentation (chaining). The M-bit is set, if more fragments, that also belong to the currently transmitted chain, will follow.

Various supervision requests and responses are defined to control the properties of the connection. The RESYNCH request allows resetting the sequence number, if devices loose synchronization. This approach slightly differs to the one in IFX-I2C, since a special frame type is defined for supervision requests and the mechanism is placed in another layer. Next, the ABORT request allows canceling an ongoing chain transmission. The SWR request is indended to trigger a software reset of the device. The CIP request allows the host controller to read out the supported communication interface parameters, e.g. maximum frame size or I2C frequency. The IFS request allows the devices to adjust their maximum frame size, depending on the used hardware interface modules and drivers. Various other supervision requests are defined, but are out of scope for our evaluation.

To support multiple logical channels, an additional header byte, the NAD (Node Address Byte), contains two fields to specify the source and destination address of a logical channel. Since the protocol stack is typically integrated in an operating system and not in an user application, this feature can be used to address multiple endpoints (e.g. applications, drivers, and kernel modules) inside the operating system. For certain physical layers, such as SPI, device addressing mechanisms can be replaced. In that case, it is possible to reduce the number of required GPIO lines for SPI by sharing a single slave select line for multiple devices.

3) Application Layer: As in IFX-I2C, interfaces between layers, as well as interfaces to the application layer and drivers are left to the implementation. This protocol is typically implemented as part of Java Card operating systems, but also native C implementations exist for bare metal applications. In case of Java Card OS integration, the application API is standardized by the GlobalPlatform Card specification.

C. Microchip cryptoauthlib

The I2C protocol stack included in Microchip's cryptoauthlib [6] has a compact and efficient design. This is achieved by including only the necessary features. The design is based on a single layer, the data link layer, which utilizes a simple request-response scheme with CRC protected frames. Whenever an I2C frame is received on the SE, the request is processed, and a result, a data frame or an error indication, is available to the host controller.

A special optimization is used to handle lost packages without having sequence numbers and acknowledgments in the data link layer. The acknowledgments in the I2C physical layer are used to detect lost frames. This approach is limited to I2C, and requires a driver on the host controller that propagates error conditions to the protocol stack. Important to note, if an I2C master reads from a slave, only the address is acknowledged by the slave. To continue reading, the master acknowledges the remaining bytes. Therefore, only the loss of entire frames can be detected with this approach when reading from the slave. This provides, along with the CRC checksums, reliability in the protocol.

Packet fragmentation is not supported. Instead, fragmentation is implemented on the application layer, if required. For simplicity, this protocol does not provide logical channels. Regarding timeout handling, polling is supported, as well as fixed wait times. For the latter, execution times of all supported commands on application level have been measured and stored in the host controller. Finally, the connection interface parameters are fixed in the protocol implementation, and are not negotiated during protocol initialization.

III. Design

The protocol stacks depicted in the previous section are either complex to implement (IFX-I2C and GlobalPlatform) or lack certain features (Microchip). With our design we intend to keep the implementation effort low, while still supporting necessary features, such as reliability, packet fragmentation, and connection interface parameter configuration. Therefore, we propose TinyI2C, a lightweight protocol stack for I2C, implementable with low memory footprint. Figure 4 depicts the design, consisting of two layers, the physical layer and the data link layer.

A. Physical Layer

The physical layer provides a set of I2C registers. The DATA register is used to relay packets to the upper layer. Transmitting a data packet consists of writing the total packet length to the DATA_LEN register and the packet to the DATA register. Packet fragmentation is implicitly integrated in the design. Data packet fragments can be written in arbitrary length to the DATA register. In addition, an abort and re-synchronization mechanism is implicitly included. Every time the DATA_LEN register is written, a new transmission is assumed, resetting the layer. Optional registers can be defined, e.g. for reading supported connection interface parameters, or triggering actions such as a software reset or power-saving mode. The physical layer provides an unreliable channel with packet fragmentation, similar to the IP protocol in the Internet.

B. Data Link Layer

The data link layer adds reliability to the communication protocol. This is achieved by using CRC checksums to detect packet errors and adding sequence numbers to detect lost and duplicate packets. The protocol distinguishes between data and control packets. Data packets are exchanged via the DATA register and control packets via the DATA_ACK register. Each transmitted packet must be acknowledged by the receiver with a control packet. Depending on the use case, the window size can be adjusted. An 8-bit field is dedicated for the sequence number, resulting in 2^8 possible values. To detect duplicate packets in case acknowledgments are lost, the maximum window size is limited to half of it, 2^4. For sake of simplicity, sequence numbers start with an initial value of zero at protocol initialization. In case peers lose synchronization, an interface reset is required. Even though there are two dedicated header bytes for the packet's payload length, the frame length is limited by the mathematical properties of used CRC checksum, in practice.

C. Additional Requirements

A major design goal is to create a symmetric protocol where the data link layer implementation can be shared between host controller and HSM. The protocol is not limited to request-response messaging schemes, it can be used for data streaming applications as well. For instance, it can be used in combination with Remote-Procedure-Call (RPC) based solutions, such as Google's gRPC, or as proposed in [5], to create distributed applications, which perform data streaming.

The protocol does not make any assumptions about the used physical communication interface. Any serial interface can be used as basis. Further, the protocol does not depend on interface or driver specific features, such as the acknowledgment on I2C.

IV. Proof-of-Concept

To show feasibility, we implemented a proof of concept in the C programming language, for deployment on resource-constraint micro-controllers. As in other protocol specifications, the interfaces between the layers are left to the implementation. We chose to implement the layers with asynchronous interfaces. This implies that there is no need for a multi-tasking operating system since transactions are non-blocking. Figure 5 depicts the interfaces and events for each layer including interface functions and event callbacks. Each layer gets a respective context structure, containing buffers and variables. This allows running multiple instances of the protocol stack in one device or process. In addition, we use only static memory-management to support especially resource-constraint systems.

The physical layer implementation differs for I2C master and slave devices. On the master side, a mechanism to read and write registers is implemented. On the slave side, buffers

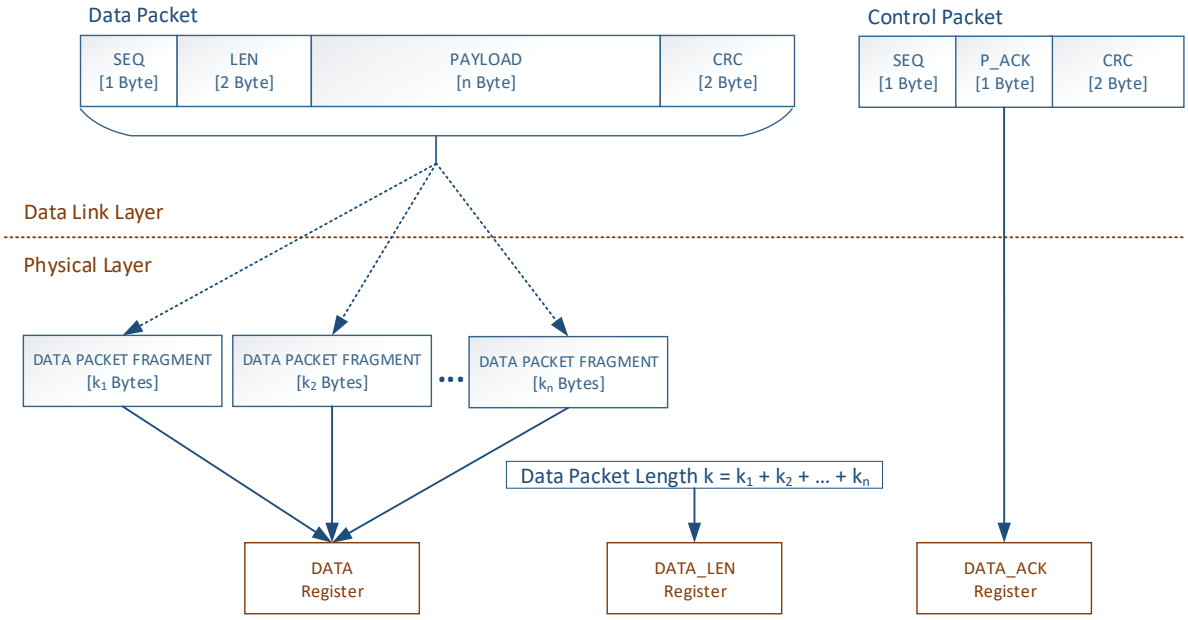


Fig. 4. TinyI2C Protocol Stack Design

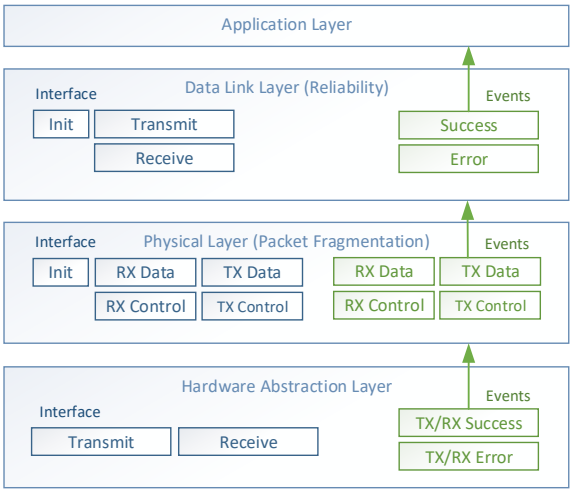


Fig. 5. TinyI2C Asynchronous Interface Implementation

and event handlers to provide these register contents are implemented. In both cases, the physical layer interface is identical. The code of the data link layer is shared for both peers, to reduce the amount of code to maintain and simplify the design.

V. Evaluation

We evaluate our concept by comparing protocol properties to existing state-of-the-art solutions.

A. Physical Layer

1) Type of Interaction: The depicted protocols have different data transmission mechanisms at the physical layer. IFX-I2C and TinyI2C use a set of I2C registers at this layer. This concept is common for sensors with I2C interface. GlobalPlatform and Microchip use a request-response state-machine. Such a state-machine is slightly easier to implement, but the communication scheme is limited to a request-response scheme. To support packet streaming, the register-based variant is required. Also, registers can be added, if additional requirements appear.

2) Timeout Handling: The protocols also differ in error handling, especially in regards to operation timeout. This timeout is also referred as Block-Wait-Time (BWT) in the GlobalPlatform specification and as Polling-Timeout in IFX-I2C. All protocols support the Late-Acknowledge Algorithm for polling until the I2C slave is ready on the physical layer. The GlobalPlatform protocol defines an additional mechanism to extend the maximum waiting time for the host controller. When a slave device needs more time to process a command, a Wait-Time-Extension (WTX) supervision request is sent to the hosting device. This mechanism is adopted from the ISO 7816 standard for classic chip cards. There it was necessary to distinguish the cases where the chip card was pulled out of the card reader and the case where the card is still processing. This does not apply for I2C devices and results in unnecessary complexity. Using this mechanism allows infinite timeouts at protocol level, while the HSM is sending WTX. This shifts the responsibility to implement timeouts towards the application layer.

TABLE I
COMPARISON OF PROTOCOL STACK FEATURES

		IFX-I2C	GlobalPlatform	Microchip	TinyI2C
Physical Layer	Interaction	Registers	Request-Response	Request-Response	Registers
	CIP Configuration	Register	Supervision Request (DL Layer)	Static Values	Register
	Timeout Handling	Polling	Polling/Interrupt	Fixed Wait Times	Polling/Interrupt
Data Link Layer	Interaction	Request-Response	Request-Response	Request-Response	Streaming
	Sequence Numbers	2 bit	1 bit	n/a	8 bit
	CRC Checksums	Yes	Yes	Yes	Yes
	Header Size	5	6	4	5
	Symmetry	Yes	No	No	Yes
Any Layer	Fragmentation	Yes (above Reliability)	Yes (above Reliability)	n/a	Yes (below Reliability)
	Logical Channels	Yes (4 bit)	Yes (8 bit)	n/a	n/a

3) Optional Interrupt Line: The GlobalPlatform protocol defines an optional feature, an interrupt line (GPIO), as alternative to polling. This feature can reduce energy consumption compared to polling. The potential savings depend on the otherwise used polling intervals. Even though not defined as a mandatory requirement, this feature can also be implemented for TinyI2C. To provide maximum flexibility, the choice of the used mechanism it is left to the implementation.

4) Connection Interface Parameters: For the evaluated protocols, configuration of the connection interface parameters is handled in two different approaches. Microchip's cryptoauthlib has hard-coded CIP values in the protocol implementation for each supported product. This brings two advantages; first, no CIP values have to be exchanged when initiating the protocol. Second, no additional provisioning is required, since the system integrator must anyway place the protocol stack and support libraries onto the hosting device to use the HSM.

GlobalPlatform defines a supervision request in the data link layer, the CIP request to allow the hosting device to retrieve the HSM's supported CIP values. This leads to additional communication when initiating the connection, nevertheless this information could be cached on the hosting device after the first protocol initialization. On one hand since, the introduction of the CIP structure leads to additional complexity of the Standard and for implementations, on the other hand, this increases the interchangeability of implementations of different vendors. In practice, this advantage is neglectable, unless the entire application API of the HSM is standardized, as depicted in [2]. Only then, the HSM can be exchanged to a product from a different vendor without altering the host controller's software.

TinyI2C uses a register-based approach for the physical layer, as IFX-I2C. In both protocols, registers are used to store the CIP values on the HSM. The registers are optional and can be skipped if vendors use static configurations, as Microchip. Unlike GlobalPlatform and IFX-I2C, the format of these registers is not specified in the TinyI2C specification, since they belong to the physical layer and depend on the used hardware, interfaces, and application.

B. Data Link Layer

1) Frame Header Encoding: Compared to the GlobalPlatform protocol, TinyI2C aims to be a symmetric protocol at the data link layer. This allows using the same implementation for hosting device and HSM. GlobalPlatform's concept of supervision requests requires implementation of request and response handlers, resulting in asymmetry.

A major difference to the GlobalPlatform and IFX-I2C protocols is that TinyI2C does not carry acknowledgments in data packets. Instead, the I2C register DATA_ACK is used. This leads to immediate acknowledgments, because the data link layer does not wait for response data from the application layer. In addition, this design simplifies implementations, since different buffers are used for data and control frames.

2) Reliability: All evaluated protocols use CRC-16-CCITT checksums according to [9]. This can protect data up to a length of 4093 bytes. In regard of sequence numbers, the protocols differ. GlobalPlatform uses a 1-bit sequence number, since a request-response communication scheme is used on the data link layer. The IFX-I2C protocol defines a 2-bit sequence number, and allows adjusting the window size between 1 and 2. In practice, the window size is fixed to 1 in the official implementation and this feature remains unused. Microchip's cryptoauthlib I2C protocol does not use sequence numbers. This is possible, because only a request-response scheme is used in combination with the acknowledgments in the I2C physical layer.

Even though IFX-I2C supports the concept of window size in the data link layer, streaming support is not available due to the implementation of the higher layers. The official implementation limits the communication to an APDU-based request-response scheme. If the data link layer design is considered solely, then limited streaming support with window size of 2 would be possible.

TinyI2C assumes an unreliable channel and does not dependent on any special feature of the underlying physical layer. It provides an 8-bit sequence number field. Choosing and adjusting the windows size is up to the implementation. This allows on the one hand simple implementations for a window size fixed to 1, and enables flexibility, if an application requires packet streaming.

Unlike the other protocols, TinyI2C does not specify a mechanism for re-synchronization. The reason is that transmission errors are not an issue in practice, if the PCB is properly designed. Unrecoverable errors are typically related to faulty protocol implementations or PCB design flaws. If the protocol

TABLE II
COMPARISON OF PROTOCOL STACK CODE SIZE

Protocol Stack	Code Size
Infineon I2C Protocol Stack	5.5 kB
GlobalPlatform APDU Transfer over I2C	2.1 kB
Microchip cryptoauthlib I2C	0.8 kB
TinyI2C Protocol	1.8 kB

stack of the HSM is in an invalid state, the hosting device might not be able to successfully transmit a software reset request. It is recommended to implement an interrupt line or power-on-reset mechanism for the HSM to recover in such cases.

3) Logical Channels: The GlobalPlatform and IFX-I2C protocols specify bits in the protocol header to support multiple logical channels. Placing this multiplexing functionality at protocol level results in a more complex application interface. As in Microchip's cryptoauthlib, we chose not to support logical channels, to keep the protocol components as simple as possible. Since a typical HSM runs only a single application, this feature remains unused in most cases, and only adds unnecessary complexity. Certain Java Card based HSMs provide an additional multiplexing mechanism on application level to select multiple Java Card Applets simultaneously. TinyI2C can also be used with logical channels implemented on application level, if required, but the application interface remains simple.

4) Packet Fragmentation: IFX-I2C and GlobalPlatform define a packet fragmentation mechanism in a (sub-)layer above the reliability-functionality. This implies that each fragment must be acknowledged by the receiver. For TinyI2C, we placed packet fragmentation below the reliability-functionality in the protocol stack, as in the well-known Internet protocol stack, where fragmentation (IP) is below reliability (TCP). This implies that only one acknowledgment for an entire packet is required, regardless the number of fragments. This approach reduces the number of transactions, and therefore improves speed. This improvement is reflected, whenever packets with multiple fragments are transmitted, which depends on the application.

5) Code Size Estimation: To estimate the complexity of the evaluated protocols, we depict the code size in Table II. Code size depends on the used compiler, preprocessor settings, target system, compiler optimizations, and other factors. This table only intended to give a rough estimator. Values are obtained from the GNU arm-gcc compiler, with size optimization. For IFX-I2C and Microchip protocols, official reference implementations are taken. For GP and TinyI2C, the authors implementations are taken into account. Our TinyI2C implementation does not have the smallest code size, but provides additional features compared to other solutions.

6) Feature Comparison: Finally, we provide a summary of all discussed features in Table I.

VI. CONCLUSION

In this paper, we proposed TinyI2C, a protocol stack indented to connect HSMs to their hosting devices. Even though HSMs were considered as primary use-case, the protocol can be adapted for other use-cases as well. We defined the core features, reliability and packet fragmentation, in an efficient way, while leaving flexibility for vendors to extend the protocol to their needs. The evaluation shows, TinyI2C supports most features and allows simple implementations, as depicted in the code size comparison. The protocol provides a symmetric data link layer, where code can be shared between both peers, and provides packet streaming support. Latter allows using the protocol in combiniation with RPC-based frameworks, such as Google's gRPC. Finally, the protocol can be adapted to other serial interfaces and extended with new applications, platforms, and vendor specific features due to its flexible design.

REFERENCES

[1] APDU Transport over SPI/I2C v1.0. Standard, GlobalPlatform, Inc., January 2020.

[2] G. Arfaoui, S. Gharout, and J. Traoré. Trusted execution environments: A look under the hood. In *2014 2nd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, pages 259–266, 2014.

[3] Bluetooth SIG. *Bluetooth Specification Version 4.2*. 2014.

[4] V. Carvalho and F. L. Kastensmidt. Enhancing i2c robustness to soft errors. In *2017 IEEE 8th Latin American Symposium on Circuits Systems (LASCAS)*, pages 1–4, 2017.

[5] T. Fischer, C. Lesjak, D. Pirker, and C. Steger. Rpc based framework for partitioning iot security software for trusted execution environments. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0430–0435, 2019.

[6] GitHub Repository. CryptoAuthLib - Microchip CryptoAuthentication Library. https://github.com/MicrochipTech/cryptoauthlib. [Online; accessed 2020-04-20].

[7] D. Gruss, M. Lipp, M. Schwarz, D. Genkin, J. Juffinger, S. O'Connell, W. Schoechl, and Y. Yarom. Another flip in the wall of rowhammer defenses. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 245–261, May 2018.

[8] IETF. *RFC793 - Transmission Control Protocol*. tools.ietf.org, 1981.

[9] ISO 13239:2002, Information technology — Telecommunications and information exchange between systems — High-level data link control (HDLC) procedures. Standard, International Organization for Standardization, Geneva, CH, 202.

[10] ISO 7816-3:2006, Identification cards — Integrated circuit cards — Part 3: Cards with contacts — Electrical interface and transmission protocols. Standard, International Organization for Standardization, Geneva, CH, March 2006.

[11] R. S. S. Kumari and C. Gayathri. Interfacing of mems motion sensor with fpga using i2c protocol. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–5, 2017.

[12] V. Lipovac. Practical analysis of the stability of tcp. In *Proceedings of the 2001 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference. (Cat. No.01TH8568)*, volume 1, pages 277–280 vol.1, 2001.

[13] Marcus Janke, Dr. Peter Laackmann. Attacks on Embedded Devices. Embedded World Conference Nurenberg, 2016.

[14] Specification. Infineon I2C Protocol Stack v2.02. https://github.com/Infineon/optiga-trust-m/blob/master/documents/Infineon_I2C_Protocol_v2.02.pdf. [Online; accessed 2020-04-20].

[15] Trusted Computing Group. TPM 2.0 Library Specification. https://trustedcomputinggroup.org/resource/tpm-library-specification/. [Online; accessed 2020-04-20].

The Local Rényi Entropy Based Shrinkage Algorithm for Sparse TFD Reconstruction

Vedran Jurdana
Department of Automation and Electronics
University of Rijeka, Faculty of Engineering
Rijeka, Croatia
vjurdana@riteh.hr

Ivan Volaric
Department of Automation and Electronics
University of Rijeka, Faculty of Engineering
Rijeka, Croatia
ivolaric@riteh.hr

Victor Sucic
Department of Automation and Electronics
University of Rijeka, Faculty of Engineering
Rijeka, Croatia
vsucic@riteh.hr

***Abstract*—Observing a non-stationary signal with the time and frequency representation being mutually exclusive often does not provide enough information. Thus, the joint time-frequency distribution (TFD) is used as a convenient and powerful tool for analysis of such signals. Although TFD overcomes many signal representation limitations, it also introduces additional challenges. The removal of artefacts, also called the cross-terms, while maintaining a high concentration of the signal components (auto-terms) is the main problem of the time-frequency (TF) signal analysis. Among different approaches of solving this problem, in this paper we are investigating the advantages of the TFD sparsity, that is, the fact that the energy is accumulated around the instantaneous frequency law of the signal components. In this paper, we present a sparse TFD reconstruction algorithm based on the iterative shrinkage algorithm. The shrinkage is performed independently for each TFD time- and frequency-slice by taking advantage obtained from the short-term and the narrow-band Rényi entropy. Using the TFD concentration measure and reconstruction algorithm execution time, the obtained results have been compared to the state-of-the-art sparse reconstruction algorithms.**

***Index Terms*—time-frequency distribution; short-term Rényi entropy; narrow-band Rényi entropy; sparse signal reconstruction**

I. Introduction

Complete understanding of the signal information is important, but sometimes it is difficult to achieve. In practice, signals are often nonstationary, thus limiting application of the Fourier transform (FT). In that case, the signal spectrum, $S(f)$, does not show time-dependent frequency variations, an information which is crucial for the nonstationary signal characterization. Spectral representation of such signals requires the mutually exclusive time and frequency variables [1], [2]. To overcome these limitations, a joint time-frequency distribution (TFD) of the signal has been introduced [1], [2]. Although TFDs provide a powerful set of tools for the signal analysis, they also have opened new challenges. Most TFD calculation methods result with the additional energy clusters which can be classified as interferences, artefacts or cross-terms, limiting the application of the time-frequency (TF) signal analysis. Signal components in the ideal TFD should be delta functions, following the instantaneous frequency law of the signal components (auto-terms), without producing the unwanted cross-terms between them. Different approaches have been introduced in order to achieve a TFD approximating this ideal case [1], [2].

Over the last years, we have witnessed an emergence of the methods for the signal reconstruction from a small sub-set of samples, introducing a new topic in the signal processing often referred to as the signal sparsity and the compressive sensing (CS). This signal property has been used in many applications in popular fields: medicine, radar, communications, etc. [3]–[7]. Taking the advantage of the TFD sparsity, the time-frequency domain can be combined with the CS based methods [5], [7]–[10]. The idea behind the CS gives an opportunity to suppress the cross-terms without the significant auto-terms resolution loss by using a small number of signal samples in the alternative domain called the ambiguity function (AF), calculated as a 2D FT transformation of the TFD [8], [10]. Since the auto-terms are less oscillatory in the TF domain, they are positioned near the AF domain origin. These samples are processed by a sparse reconstruction algorithm with the goal of achieving a high resolution sparse TFD. The sparse reconstruction algorithm solves the unconstrained optimization problem, with the objective function emphasizing the TFD sparsity level [11], [13]. In this paper, we present a sparse TFD reconstruction algorithm which uses the information about the instantaneous number of signal components, obtained from the short-term Rényi entropy [14], [15] and the here-proposed narrow-band Rényi entropy.

The organization of the paper is as follows. Section II gives a short overview of the quadratic class of TFDs (QTFD), along with the short-term and the narrow-band Rényi entropy, while in Section III the sparse TFD reconstruction is introduced and the proposed sparse TFD reconstruction

This work has been fully supported by the University of Rijeka under the project number UNIRI-TEHNIC-18-67.

algorithm is presented. Section IV presents the simulation results of the proposed method, which are compared to state-of-the-art sparse reconstruction algorithms in terms of the TFD concentration measure and the reconstruction algorithm execution time. Section V gives the conclusion of the paper.

II. Time-frequency distribution

A. Quadratic Time-Frequency Distributions

Ville defined the most commonly used QTFD by modifying the Wigner Distribution by replacing a real signal, $s(t)$, with its analytic associate, $z(t)$. This QTFD is referred to as the Wigner-Ville Distribution (WVD), and is defined as [2]:

$$W_z(t,f) = \int_{-\infty}^{\infty} K_z(t,\tau) e^{-j2\pi f\tau} d\tau, \tag{1}$$

where the instantaneous autocorrelation function (IAF), $K_z(t,\tau)$, is defined as:

$$K_z(t,\tau) = z\left(t+\frac{\tau}{2}\right) z^*\left(t-\frac{\tau}{2}\right). \tag{2}$$

The WVD provides a very good representation for a signal with a single linear frequency modulated (LFM) component. However, in case of the multi-component signal:

$$z(t) = \sum_{i=1}^{N_c} a_i(t) e^{j\phi_i(t)}, \tag{3}$$

where N_c is the number of components, $a_i(t)$ is the instantaneous amplitude and $\phi_i(t)$ is the instantaneous phase of the i-th component, the WVD of a signal contains the WVDs of each component and the cross-WVDs between each pair of the components, resulting with the useful components, named the auto-terms, and the cross-terms [2]. The multi-component IAF is given by:

$$\begin{aligned} K_z(t,\tau) = &\underbrace{\sum_{i=1}^{N_c} z_i\left(t+\frac{\tau}{2}\right) z_i^*\left(t-\frac{\tau}{2}\right)}_{\text{auto-terms}} + \\ &+ \underbrace{\sum_{i=1}^{N_c}\left[z_i\left(t+\frac{\tau}{2}\right)\sum_{j=1,j\neq i}^{N_c} z_j^*\left(t-\frac{\tau}{2}\right)\right]}_{\text{cross-terms}}, \end{aligned} \tag{4}$$

where the cross-terms appear due to the superposition rule, resulting in the WVD:

$$W(t,f) = \underbrace{\sum_{i=1}^{N_c} W_i(t,f)}_{\text{auto-terms}} + \underbrace{2\sum_{i=1}^{N_c}\sum_{j=1,j\neq i}^{N_c} \mathrm{Re}\{W_{i,j}(t,f)\}}_{\text{cross-terms}}. \tag{5}$$

In order to achieve adequate cross-terms suppression, the WVD, having the best auto-term resolution and the worst proneness for the cross-terms, has been improved by adding a low-pass filter leading to a definition of the filtered QTFD class:

$$\rho_z(t,f) = W_z(t,f) \underset{t}{*}\underset{f}{*} \gamma(t,f), \tag{6}$$

where $\gamma(t,f)$ is the low-pass filter kernel [2]. To avoid a double convolution in the TF domain, the kernel design is usually performed in the AF, a 2D half inverse, half forward FT of the WVD, calculated as:

$$\mathcal{A}_z(\nu,\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_z(t,f) e^{j2\pi(f\tau-\nu t)} dt df, \tag{7}$$

where (6) is replaced with:

$$\mathcal{A}_z(\nu,\tau) = \mathcal{A}_z(\nu,\tau) g(\nu,\tau), \tag{8}$$

where $g(\nu,\tau)$ is the AF filter kernel. The non-oscillating auto-terms are positioned around the AF domain origin, while the highly-oscillating cross-terms are positioned through the rest of the domain. Thus, the idea is to design a low-pass filter that will take as many auto-terms as possible, and in the same time, filter out as much of the cross-terms. In practice, the cross-terms get filtered out at the expense of the auto-terms resolution, and the appropriate kernel design involves a trade-off between these two [2].

B. The Short-Term Rényi Entropy

Williams, Brown and Hero employed the generalized Rényi entropies [16] in order to measure the amount of information within the TF plane. The Rényi entropy is defined as:

$$R_z^\alpha(\rho_z(t,f)) = \frac{1}{1-\alpha}\log_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \rho_z^\alpha(t,f) dt df, \tag{9}$$

where parameter $\alpha > 2$ is chosen as an odd integer value in order to cancel the cross-terms which are integrated out over the entire TF plane, that is:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_{i,j}(t,f) dt df = 0. \tag{10}$$

A method for estimating the instantaneous number of signal components based on the Rényi entropy has been proposed in [14], [15]. Instead of taking the Rényi entropy over the entire TF plane, this method observes and calculates the Rényi entropy of each time-slice t_0. The foundation for this method is in fact that by taking a sub-set of TFD samples in the vicinity of t_0, different components locally have same time durations and similar bandwidths, overcoming the global Rényi entropy inaccuracy when the components exhibit different time and frequency supports, and if one of the components is not known in advance [14], [15]. The instantaneous number of signal components, $N_{c,t}(t_0)$, is calculated as:

$$N_{c,t}(t_0) = 2^{R_z^\alpha(\Delta^t\rho_z(t,f)) - R_z^\alpha(\Delta^t\rho_{\text{ref}}(t,f))}, \tag{11}$$

where t_0 is the observed time-slice, $\rho_{ref}(t,f)$ is the TFD of the reference signal and $\rho_z(t,f)$ is the TFD of the original signal. The notation Δ^t denotes that all TFD samples are set to zero, except those in the vicinity of t_0:

$$\Delta^t\rho_z(t,f) = \begin{cases} \rho_z(t,f), & t_0-\Delta t < t < t_0+\Delta t, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where Δt is the user defined localization parameter defining the length of the observed time interval. The reference signal

is usually chosen to be a signal with a constant normalized frequency of 0.1 [14], [15]. Furthermore, $\rho_z(t,f)$ and $\rho_{\text{ref}}(t,f)$ must be calculated using the same TFD kernel [14], [15].

C. The Narrow-Band Rényi Entropy

Using the same motivation and idea as with the short-term Rényi entropy [14], [15], in this paper we propose calculation of the local Rényi entropy in short frequency intervals, as well. The number of data points in a frequency-slice, $N_{c,f}(f_0)$, is calculated analogue to (11), with the only difference that Δ^t is replaced with Δ^f:

$$\Delta^f \rho_z(t,f) = \begin{cases} \rho_z(t,f), & f_0 - \Delta f < f < f_0 + \Delta f, \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where f_0 is the observed frequency-slice, and Δf is the user defined localization parameter defining the length of the observed frequency slice. However, the main difference comparing the narrow-band Rényi entropy with the short-term Rényi entropy is in the reference signal which is chosen to be a delta function at $t = 15$ in conducted simulations.

The selected delta function, when observed in the TF plane is perfectly localized, while containing all frequencies, which is opposite to the cosine signal selected for the short-term Rényi entropy. Both local Rényi entropy approaches achieve more accurate results analysing signals with components more aligned to their reference signal, which will be also shown in Section IV. With that observation, our goal was to additionally improve the information about the number of local components by combining both approaches in the sparse reconstruction algorithm.

III. The Sparse TFD Reconstruction

A. Sparse Time-Frequency Distributions

The idea in the sparse TFD calculation is to take a small AF area around the domain origin (also referred as the CS) assuming that the selected area will not contain the cross-terms. The rest of the AF is then calculated by minimizing the sparsity level of the resulting TFD [8], [10].

The AF filtering can be rewritten as the CS in the matrix form as:

$$\mathbf{A}_z^{'}(\nu,\tau) = \phi(\nu,\tau) \odot \mathbf{A}_z(\nu,\tau), \tag{14}$$

where the operator $\odot$ denotes element-by-element matrix multiplication, $\mathbf{A}_z'(\nu,\tau)$ is the discretized AF, while $\phi(\nu,\tau)$ is the sensing matrix defining the $N_\nu' \times N_\tau'$ area Ω around the AF origin:

$$\phi(\nu,\tau) = \begin{cases} 1, & (\nu,\tau) \in \Omega, \\ 0, & \text{otherwise}. \end{cases} \tag{15}$$

The sparse TFD, $\boldsymbol{\vartheta}_z(t,f)$, is calculated as:

$$\boldsymbol{\vartheta}_z(t,f) = \boldsymbol{\psi}^H \cdot \mathbf{A}_z^{'}(\nu,\tau), \tag{16}$$

where $\boldsymbol{\psi}^H$ is the Hermitian transpose of a domain transformation matrix representing the 2D Fourier transformation equivalent to (7) [10]. The goal of the reconstruction algorithm is to find the optimal solution of (16), but since this problem does not have a unique solution, we have to introduce the regularization function, emphasizing the desirable solution properties [13]. The problem in (16) is an unconstrained optimization problem written as [13]:

$$\hat{\boldsymbol{\vartheta}}_z(t,f) = \arg\min_{\boldsymbol{\vartheta}_z(t,f)} \tfrac{1}{2}||\boldsymbol{\vartheta}_z(t,f) - \boldsymbol{\psi}^H \mathbf{A}_z^{'}(\nu,\tau)||_2^2 + \lambda c(\boldsymbol{\vartheta}_z(t,f)), \tag{17}$$

where $c(\boldsymbol{\vartheta}_z(t,f))$ is the regularization function multiplied with the regularization parameter λ. To achieve a low sparsity level of the solution, the ℓ_q-norm with $0 \leq q \leq 1$ is used [5], [7], [10], [13]. More precisely, the ℓ_0-norm and the ℓ_1-norm have been used, with the ℓ_1-norm commonly used as the regularization function due to its convexity, and the ℓ_0-norm as the best sparsity inducing function [8], [10]. However, the ℓ_0-norm minimization cannot be solved in polynomial time, thus it must be iteratively approximated with the greedy algorithms or the hard-thresholding based algorithms [17], [18].

In the proposed sparse TFD reconstruction algorithm, the ℓ_0-norm-based regularization function has been used. The optimization problem is thus, defined as [13], [17], [18]:

$$\begin{aligned} \boldsymbol{\vartheta}_z^{\ell_0}(t,f) &= \arg\min_{\boldsymbol{\vartheta}_z(t,f)} ||\boldsymbol{\vartheta}_z(t,f)||_0, \\ &||\boldsymbol{\vartheta}_z(t,f) - \boldsymbol{\psi}^H \mathbf{A}_z^{'}(\nu,\tau)||_2^2 \leq \epsilon, \end{aligned} \tag{18}$$

where ϵ defines a user-defined solution tolerance. The proximity operator has been introduced, transforming (18) into a closed-form expression [17], [18]:

$$\boldsymbol{\vartheta}_z^{\ell_0}(t,f) = \text{hard}_{\sqrt{2\lambda}}\{\boldsymbol{\vartheta}_z(t,f)\}, \tag{19}$$

where $\text{hard}_{\sqrt{2\lambda}}\{\boldsymbol{\vartheta}_z(t,f)\}$ is a hard-thresholding function defined as:

$$\text{hard}_{\sqrt{2\lambda}} = \begin{cases} \boldsymbol{\vartheta}_z(t,f), & \boldsymbol{\vartheta}_z(t,f) \geq \sqrt{2\lambda}, \\ 0, & \text{otherwise}. \end{cases} \tag{20}$$

Effective algorithm for the iterative sparse signal reconstruction on which the proposed algorithm is based is the two-step iterative shrinkage/thresholding (TwIST) algorithm:

$$\begin{aligned} &\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n+1]} = (1-\alpha)\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n-1]} + (\alpha-\beta)\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n]} + \\ &+\beta \text{hard}_{\sqrt{2\lambda}}\left\{\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n]} + \boldsymbol{\psi}^H\left(\mathbf{A}_z^{'}(\nu,\tau) - \boldsymbol{\psi}\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n]}\right)\right\}, \end{aligned} \tag{21}$$

where α and β are the user-defined parameters. The resulting sparse TFD is calculated by iterating (21) until the stopping criterion is satisfied or the user-defined maximum number of iterations, N_{it}, is reached.

B. The Proposed Time- and Frequency-Slice Shrinkage Algorithm

In [19] a different way of looking at the hard-thresholding has been investigated. Instead of removing the low-valued samples based on λ, we can keep high-valued samples based on the numbers of local components, $N_{c,t}$ and $N_{c,f}$, obtained from two local entropy approaches: the short-term and the narrow-band Rényi entropy, respectively. Although taking samples around the highest local maxima seems convenient,

that approach showed faulty in some tests, since the cross-terms often have larger maxima then the auto-terms. Thus, the here-proposed algorithm is based on the assumption that the auto-terms have the largest nonnegative energy surface in a TFD time- and frequency-slice.

To begin with, all the negative TFD values are set to zero:

$$\boldsymbol{\vartheta}_z(t,f) = \text{hard}_0\{\boldsymbol{\vartheta}_z(t,f)\}. \tag{22}$$

The reason behind this is that the auto-term energy is always positive, thus all negative samples belong exclusively to the cross-terms. Second step is to find all local maxima in the current time- or frequency-slice which can be either the auto-term or the cross-term related. In the next step, we calculate the surface around each maximum. Next, the new time- or frequency-slice is generated with the samples belonging to $N_{c,t}$ or $N_{c,f}$ largest surfaces, while disregarding all other samples. The user can control the amount of transferred samples with the additional parameters: δ_t in time-slice and δ_f in frequency-slice, where $0 \leq \delta_{t,f} \leq 1$. For $\delta_{t,f} = 1$ only surface maxima will be transferred, while for $\delta_{t,f} = 0$ all surface samples will be transferred in a new time- or frequency-slice. Finally, the current TFD time- or frequency-slice is replaced with a newly generated slice and the whole process is repeated for all time- and frequency-slices. The pseudocode for this algorithm has been given in Algorithm 1.

Let us denote the argument of the hard-thresholding operator in (21) with $\boldsymbol{\varsigma}'_z(t,f)$, and the hard-thresholding result as $\boldsymbol{\varsigma}_z(t,f)$, that is:

$$\boldsymbol{\varsigma}'_z(t,f) = \left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n]} + \boldsymbol{\psi}^H\left(\mathbf{A}'_z(\nu,\tau) - \boldsymbol{\psi}\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[n]}\right), \tag{23a}$$

$$\boldsymbol{\varsigma}_z(t,f) = \text{hard}_{\sqrt{2\lambda}}\left\{\boldsymbol{\varsigma}'_z(t,f)\right\}. \tag{23b}$$

In the proposed algorithm, (23b) has been replaced with the shrinkage operator:

$$\boldsymbol{\varsigma}_z^{t,f}(t,f)^{[n+1]} = \text{shrink}_{t,f}\left\{\boldsymbol{\varsigma}'_z(t,f)\right\}, \tag{24}$$

where $\text{shrink}_{t,f}\{\cdot\}$ operator shrinks the TFD in each reconstruction algorithm iteration. The $\text{shrink}_t\{\cdot\}$ returns $\boldsymbol{\varsigma}_z^t(t,f)$ with shrunken time-slices, while the $\text{shrink}_f\{\cdot\}$ returns $\boldsymbol{\varsigma}_z^f(t,f)$ with shrunken frequency slices. In our experiments, we have combined these two TFDs in two ways. First is by averaging $\boldsymbol{\varsigma}_z^t(t,f)$ and $\boldsymbol{\varsigma}_z^f(t,f)$, that is:

$$\boldsymbol{\varsigma}_z(t,f) = p\cdot\boldsymbol{\varsigma}_z^t(t,f) + (1-p)\cdot\boldsymbol{\varsigma}_z^f(t,f), \tag{25}$$

where p is the user defined weighting parameter. The second approach keeps the sample if it is present in both $\boldsymbol{\varsigma}_z^t(t,f)$ and in $\boldsymbol{\varsigma}_z^f(t,f)$, that is:

$$\boldsymbol{\varsigma}_z(t,f) = \begin{cases} 0, & \boldsymbol{\varsigma}_z^t(t,f) \neq \boldsymbol{\varsigma}_z^f(t,f), \\ \boldsymbol{\varsigma}_z^t(t,f), & \boldsymbol{\varsigma}_z^t(t,f) = \boldsymbol{\varsigma}_z^f(t,f). \end{cases} \tag{26}$$

The full sparse TFD shrinkage algorithm has been given in Algorithm 2.

Algorithm 1 Time-/frequency-slice shrinkage algorithm

Function $fun\_Thresh(\boldsymbol{\varsigma}'_z(tf,ft),\delta,N_c)$:
$\boldsymbol{\varsigma}'_z(tf,ft) \leftarrow 0$;
Solve (22)
for $i \leftarrow 1$ **to** $length\,(\boldsymbol{\varsigma}'_z(tf,:))$ **do**
  $slice(:) \leftarrow \boldsymbol{\varsigma}'_z(i,:)$;
  $max_{ind} \leftarrow$ indexes of slice local maxima;
  for $j \leftarrow 1$ **to** $length\,(max_{ind})$ **do**
    $surf(j,3) = max_{ind}(j)$;
    $surf(j,2) \leftarrow$ local minimum to the left;
    $surf(j,1) \leftarrow$ local minimum to the right;
    $surf(j,0) \leftarrow$ sum of samples from $surf(j,2)$ to $surf(j,1)$;
  end for
  $surf \leftarrow$ ascendingly sort w.r.t to $surf(:,0)$;
  for $j \leftarrow 1$ **to** $N_c(i)$ **do**
    for $k \leftarrow surf(j,2)$ **to** $surf(j,1)$ **do**
      if $slice(k) \geq \delta \cdot surf(j,3)$ **then**
        $\boldsymbol{\varsigma}_z(i,k) = slice(k)$
      end if
    end for
  end for
end for
return $\boldsymbol{\varsigma}_z(tf,ft)$

IV. Experimental results

The here-proposed algorithm performance has been tested on three synthetic signals: the signal with $N_t = 256$ samples composed of three LFM components with different amplitudes, z_{3LFM}; the signal with $N_t = 256$ samples composed of four LFM components, z_{4LFM}; and the signal with $N_t = 128$ samples composed of one LFM and one sinusoidal FM component, $z_{\text{LFM,SIN}}$. The WVDs of the considered signals and their AFs are shown in Fig. 1. The CS-AF area has been selected by the method proposed in [10]. The reconstruction performance has been compared with the following state-of-the-art reconstruction algorithms: TwIST [11], Sparse reconstruction by separable approximation (SpaRSA) [20],

Algorithm 2 Sparse TFD reconstruction algorithm

Input: $\mathbf{A}'_z(\nu,\tau), \boldsymbol{\psi}, \boldsymbol{\psi}^H, N_{c,t}, N_{c,f}, \alpha, \beta, \epsilon, N_{it}, \delta_t, \delta_f, p.$
Output: Reconstructed sparse TFD, $\boldsymbol{\vartheta}_z^{\ell_0}(t,f)$.
$\left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[-1]}, \left[\boldsymbol{\vartheta}_z^{\ell_0}(t,f)\right]^{[0]} \leftarrow \boldsymbol{\psi}^H\mathbf{A}'_z(\nu,\tau)$;
while $((c \geq \epsilon)$ **and** $(n \leq N_{it}))$ **do**
  Solve (23a)
  $\boldsymbol{\varsigma}_z^t(t,f) \leftarrow fun\_Thresh(\boldsymbol{\varsigma}'_z(t,f),\delta_t,N_{c,t})$
  $\boldsymbol{\varsigma}_z^f(t,f) \leftarrow fun\_Thresh(\boldsymbol{\varsigma}'_z(t,f)^T,\delta_f,N_{c,f})$
  Solve (25) **or** (26)
  Solve (21);
  $c \leftarrow$ stopping criterion;
  $n \leftarrow n+1$;
end while

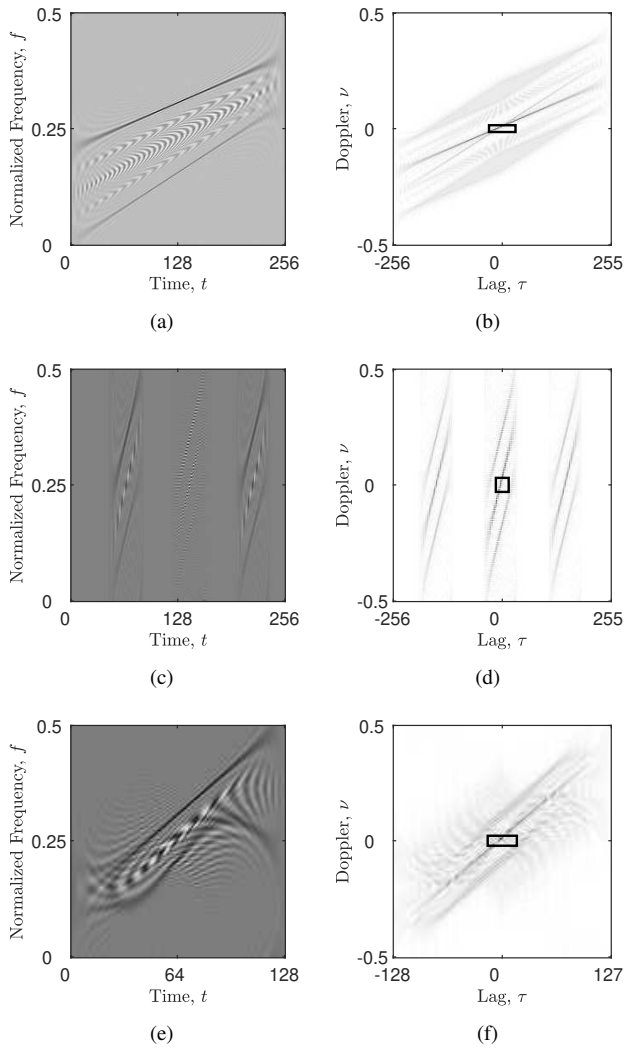


Fig. 1. Considered test signals: (a) WVD of z_{3LFM}, $M_z^S = 2.8744$; (b) AF of z_{3LFM}, $N'_\tau = 31$, $N'_\nu = 15$; (c) WVD of z_{4LFM}, $M_z^S = 1.4255$; (d) AF of z_{4LFM}, $N'_\tau = 15$, $N'_\nu = 31$; (e) WVD of $z_{\text{LFM,SIN}}$, $M_z^S = 2.7699$; (f) AF of $z_{\text{LFM,SIN}}$, $N'_\tau = 17$, $N'_\nu = 11$. The automatically selected CS-AF area, N'_τ, N'_ν, has been marked in black.

Split augmented Lagrangian shrinkage algorithm (SALSA) [21], Nestrov algorithm (NESTA) [22], and Your Augmented Lagrangian algorithm for ℓ_1 (YALL1) [12].

The algorithms have been compared based on two criteria: the algorithm execution time, and the concentration measure of the TFD, calculated as [23]:

$$M_z^S = \frac{1}{N_t N_f} \left[\sum_{t,f}^{N_t N_f} \left(\vartheta_z \left(t, f \right) \right)^{1/2} \right]^2 . \tag{27}$$

Better concentration of the corresponding TFD is proportional with the smaller M_z^S value. To minimize random behaviour of execution times, the results have been averaged over 100 algorithm runs. The simulations have been performed on a PC with the Intel Core i7-4790 @ 3.60Ghz processor and 16GB of RAM.

To achieve a reconstructed TFD with the best resolution, the consistent auto-terms and the reduced cross-terms, the shrinkage algorithm parameters have been tested in three stages, the sequence of which will be shown on the first signal example, z_{3LFM}. In first stage, for fixed parameters δ_t and δ_f, averaging of the short-term Rényi and the narrow-band Rényi entropy has been tested through parameter p. Given that all three signal components in z_{3LFM} are positioned more towards the short-term Rényi entropy reference signal, it is expected for parameter p to be closer to 1. The obtained concentration measures and the execution times are presented in Table I, with comp denoting the shrinking approach in (26). Four examples of the reconstructed TFD with a different parameter p and comp are shown in Fig. 2, with $p = 0.75$ in Fig. 2(b) selected as the best performing for z_{3LFM} given the M_z^S value and the low-amplitude auto-term preservation compromise. Next, with fixed parameter p, we optimize the parameters δ_t and δ_f, which the obtained results are shown in Table II. It can be seen in Fig. 3(a) and Fig. 3(b) that setting parameters δ_t and δ_f to lower values degrades reconstructed signal resolution.

TABLE I
SPARSE TFD CONCENTRATION MEASURE AND RECONSTRUCTION EXECUTION TIME COMPARISON OBTAINED BY VARYING THE PARAMETER p. THE BOLD VALUES INDICATE THE BEST PERFORMING AND THE FASTEST RECONSTRUCTION ALGORITHM.

	$z_{\text{3LFM}}, \delta_t = 0.88, \delta_f = 0.91$					
	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$	comp
M_z^S	0.042	0.037	0.036	0.032	0.039	**0.029**
$t[s]$	3.78	6.044	4.94	6.212	**2.624**	5.82

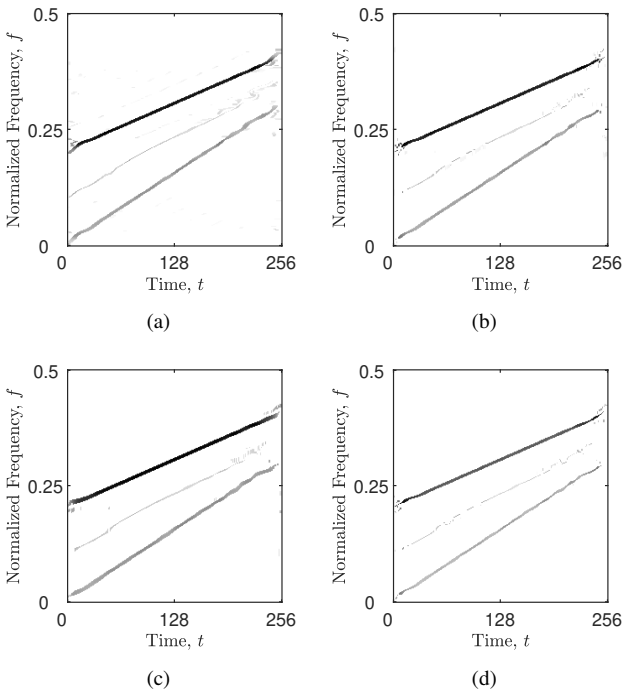


Fig. 2. Reconstructed sparse TFDs of z_{3LFM} with $\delta_t = 0.88, \delta_f = 0.91$ and: (a) $p = 0$; (b) $p = 0.75$; (c) $p = 1$; (d) comp.

TABLE II

SPARSE TFD CONCENTRATION MEASURE AND RECONSTRUCTION EXECUTION TIME COMPARISON OBTAINED BY VARYING THE PARAMETERS δ_t AND δ_f. THE BOLD VALUES INDICATE THE BEST PERFORMING AND THE FASTEST RECONSTRUCTION ALGORITHM.

	$z_{\text{3LFM}}, p = 0.75$			
	$\delta_{t,f} = 0.4$	$\delta_{t,f} = 0.6$	$\delta_{t,f} = 0.8$	$\delta_{t,f} = 1$
M_z^S	0.155	0.089	0.053	**0.007**
$t[s]$	**1.41**	2.931	4.19	37.82

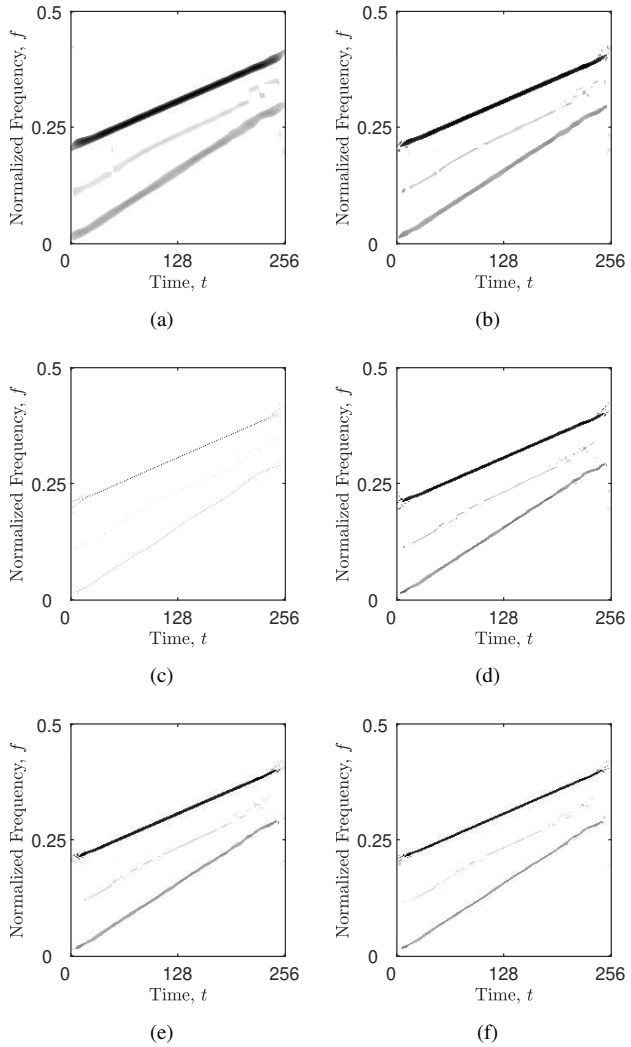


Fig. 3. Reconstructed sparse TFDs of z_{3LFM} with $p = 0.75$ and: (a) $\delta_t = \delta_f = 0.5$, $M_z^S = 0.12$, $t = 2.02$ s; (b) $\delta_t = \delta_f = 0.85$, $M_z^S = 0.044$, $t = 5.23$ s; (c) $\delta_t = \delta_f = 1$, $M_z^S = 0.007$, $t = 37.8$ s; (d) $\delta_t = 0.91, \delta_f = 0.85$, $M_z^S = 0.032$, $t = 8.897$ s; (e) $\delta_t = \delta_f = 0.91$, $M_z^S = 0.035$, $t = 3.763$ s; (f) $\delta_t = 0.95, \delta_f = 0.91$, $M_z^S = 0.027$, $t = 5.41$ s.

Although $\delta_t = \delta_f = 1$ combination, shown in Fig 3(c), produces the lowest M_z^S value, this parameter combination was excluded by visual inspection since the resulting sparse TFD is over sparse. In fact, as seen in Table II, the higher the δ_t and δ_f values are, the lower M_z^S value gets. The reason behind this is the nature of the concentration measure which favours the TFDs with less samples, and since $\delta_{t,f}$ parameters directly influence the number of samples; such behaviour was to be expected. This is why the visual inspection was the main criterion for selecting optimal $\delta_{t,f}$ parameters. Our tests have shown that setting $0.88 \leq \delta_{t,f} \leq 0.94$ is a good compromise between the auto-term resolution and component preservation. For final fine-tuning, the parameters δ_t and δ_f have been varied at small intervals. The result examples are shown in Fig. 3(d), Fig. 3(e) and Fig. 3(f) which showed that by further adjusting the parameters δ_t and δ_f, a user can mainly improve the algorithm execution time, with some minor change in the signal resolution and M_z^S value.

With the parameters set to $p = 0.75$, $\delta_t = \delta_f = 0.91$ for the signal example z_{3LFM}, the proposed Rényi entropy based algorithm has been compared with the state-of-the-art reconstruction algorithms which obtained numerical results are presented in Table III and shown in Fig. 4. Among the tested algorithms, the YALL1 algorithm has the smallest M_z^S value, followed by the here-proposed algorithm. However, by visual inspection, one might favour the here-proposed algorithm over the YALL1 because of the component concentration inconsistency. On the other hand, the resulting sparse TFD of the SALSA algorithm showed the poorest concentration, visually and according to the concentration measure. In terms of execution time, algorithms with the best concentration (the YALL1 algorithm and the here-proposed algorithm) proved as the slowest executing algorithms.

As an additional example, the proposed algorithm has been additionally tested on the synthetic signal, z_{4LFM}. Considering the position of the components in this signal, which are more aligned with the narrow-band Rényi entropy reference signal, lower values of the parameter p have been more suitable. The obtained concentration measures and the execution times are presented in Table IV. With further optimization, final reconstructed TFD, with the parameters set to $p = 0.15$, $\delta_t = 0.94, \delta_f = 0.91$, is shown in Fig. 5 alongside the state-of-the-art algorithms.

For the signal with various IF slopes, such as $z_{\text{LFM,SIN}}$, our experiments, summarized in Table IV, showed that the best results are obtained using (26) in the shrinking process. Final reconstructed TFD, with the parameters set to $\delta_t = \delta_f = 0.93$, is shown in Fig. 6, alongside the state-of-the-art algorithms. Finally, the comparison between the proposed Rényi entropy based reconstruction algorithm with the state-of-the-art reconstruction algorithms is presented in Table V.

TABLE III

COMPARISON OF THE SPARSE TFD CONCENTRATION MEASURES AND THE RECONSTRUCTION ALGORITHM EXECUTION TIMES. THE BOLD VALUES INDICATE THE BEST PERFORMING AND THE FASTEST RECONSTRUCTION ALGORITHM.

	z_{3LFM}					
	Rényi	TwIST	SpaRSA	SALSA	NESTA	YALL1
M_z^S	0.035	0.043	0.104	0.191	0.118	**0.014**
$t[s]$	3.763	0.286	**0.081**	0.161	2.329	3.368

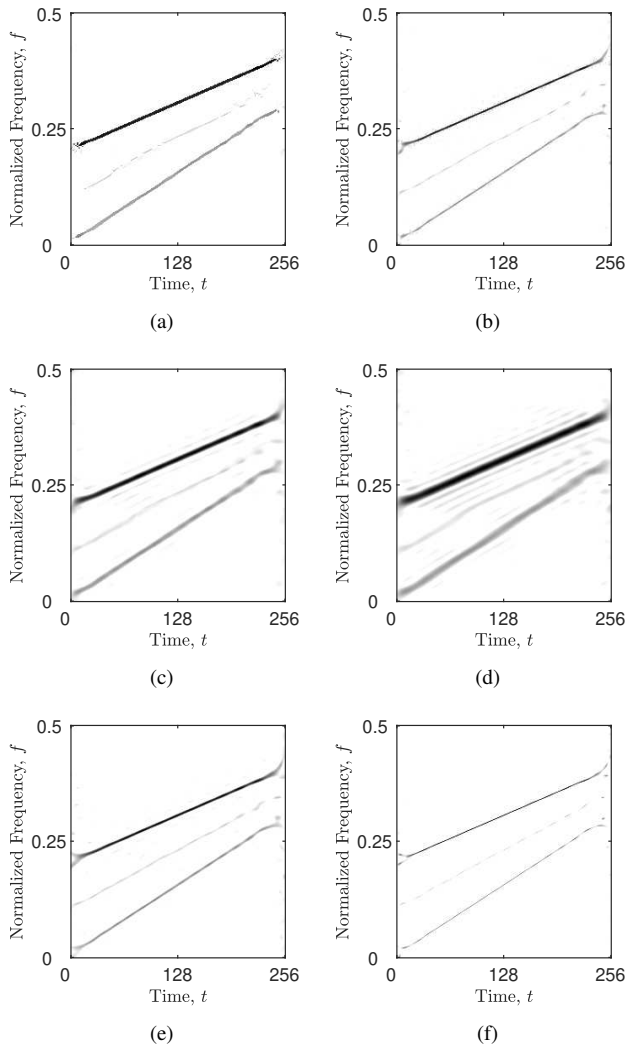


Fig. 4. Reconstructed sparse TFDs of z_{3LFM} with: (a) the proposed Rényi entropy based algorithm, $p = 0.75$, $\delta_t = \delta_f = 0.91$; (b) the TwIST algorithm; (c) the SpaRSA algorithm; (d) the SALSA algorithm; (e) the NESTA algorithm; (f) the YALL1 algorithm.

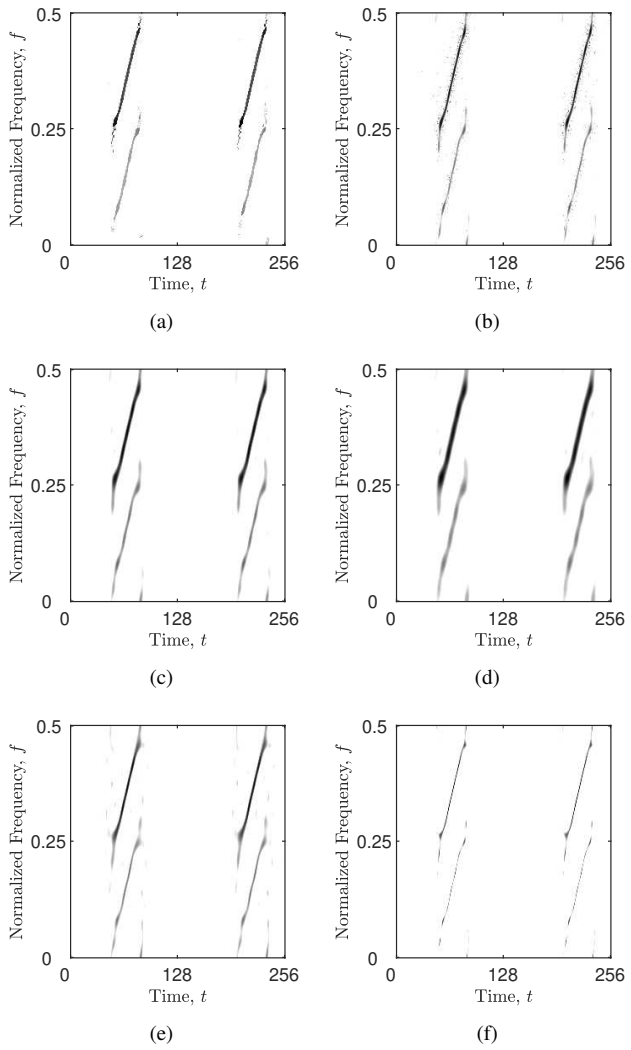


Fig. 5. Reconstructed sparse TFD of z_{4LFM} with: (a) the proposed Rényi entropy based algorithm, $p = 0.15$, $\delta_t = 0.94, \delta_f = 0.91$; (b) the TwIST algorithm; (c) the SpaRSA algorithm; (d) the SALSA algorithm; (e) the NESTA algorithm; (f) the YALL1 algorithm.

TABLE IV

SPARSE TFD CONCENTRATION MEASURE AND RECONSTRUCTION EXECUTION TIME COMPARISON OBTAINED BY VARYING THE PARAMETER p. THE BOLD VALUES INDICATE THE BEST PERFORMING AND THE FASTEST RECONSTRUCTION ALGORITHM.

	$z_{\text{4LFM}}, \delta_t = \delta_f = 0.92$					
	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$	comp
M_z^S	0.026	**0.018**	0.024	0.026	0.022	0.019
$t[s]$	**3.179**	4.927	7.722	13.532	3.278	6.811
	$z_{\text{LFM,SIN}}, \delta_t = \delta_f = 0.92$					
	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$	comp
M_z^S	0.07	0.037	0.036	0.047	0.062	**0.032**
$t[s]$	**2.192**	11.351	9.669	3.283	3.819	6.12

For both signal examples the results show an improvement in the concentration measure, ranking this algorithm right next to the YALL1.

TABLE V

COMPARISON OF THE SPARSE TFD CONCENTRATION MEASURES AND THE RECONSTRUCTION ALGORITHM EXECUTION TIMES. THE BOLD VALUES INDICATE THE BEST PERFORMING AND THE FASTEST RECONSTRUCTION ALGORITHM.

	z_{4LFM}, $p = 0.15$, $\delta_t = 0.94, \delta_f = 0.91$					
	Rényi	TwIST	SpaRSA	SALSA	NESTA	YALL1
M_z^S	0.018	0.043	0.047	0.082	0.161	**0.013**
$t[s]$	7.897	**0.133**	0.182	0.265	1.412	3.783
	$z_{\text{LFM,SIN}}$, comp, $\delta_t = \delta_f = 0.93$					
	Rényi	TwIST	SpaRSA	SALSA	NESTA	YALL1
M_z^S	0.029	0.069	0.095	0.139	0.197	**0.018**
$t[s]$	2.88	**0.042**	0.043	0.078	0.438	1.083

CONCLUSION

Application of the CS based methods to the TFD opens new possibilities of their improvement. In order to achieve a sparse TFD, the minimization problem must be solved with

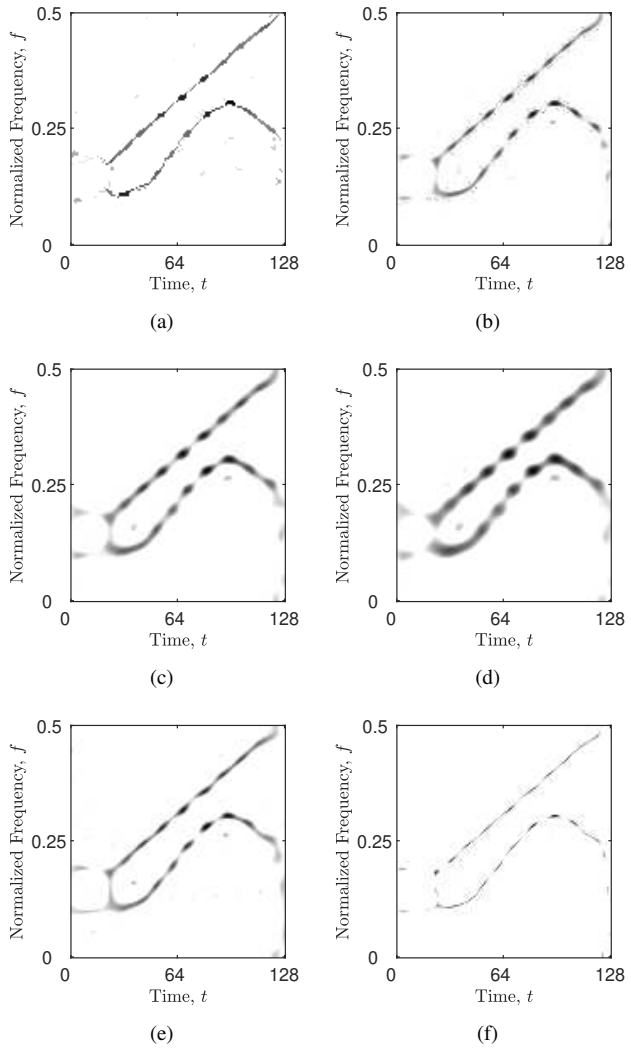


Fig. 6. Reconstructed sparse TFD of $z_{\text{LFM,SIN}}$ with: (a) the proposed Rényi entropy based algorithm, comp, $\delta_t = \delta_f = 0.93$; (b) the TwIST algorithm; (c) the SpaRSA algorithm; (d) the SALSA algorithm; (e) the NESTA algorithm; (f) the YALL1 algorithm.

an objective function emphasizing the sparsity level of the solution. In this paper, a sparse TFD reconstruction algorithm based on the iterative shrinkage algorithm has been presented. Its structure is based on the TwIST algorithm combined with the information about the local number of signal components obtained from the short-term and the narrow-band Rényi entropy. The shrinkage is performed independently for each TFD time- and frequency-slice by setting all samples to zero, except those belonging to the N_c largest surfaces. The presented results show that the here-proposed algorithm achieves competitive results with the state-of-the-art sparse reconstruction algorithms, in terms of the TFD concentration measure and the algorithm execution time, utilizing the benefits of both local Rényi entropy approaches for different signal component variations. Future research will be in the direction of testing the utility of the algorithm on noisy signals in real-life application.

REFERENCES

[1] B. Boashash, "Time-frequency signal analysis", in: S. Haykin (Ed.), Advances in Spectrum Analysis and Array Processing, vol. 1, Prentice-Hall, Englewood Cliffs, NJ, 1991, pp. 418-517.

[2] B. Boashash, Time-Frequency Signal Analysis and Processing, A Comprehensive Reference, 2nd ed., Elsevier, 2016.

[3] S. S. Vasanawala, M. T. Alley, B. A. Hargreaves, R. A. Barth, J. M. Pauly, and M. Lustig, "Improved pediatric MR imaging with compressed sensing," Radiology, vol. 256, no. 2, pp. 607–616, 2010.

[4] O. Barkan, J. Weill, A. Averbuch, and S. Dekel, "Adaptive compressed tomography sensing," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, June 2013, pp. 2195–2202.

[5] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hmlinen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," NeuroImage, vol. 70, pp. 410 – 422, 2013.

[6] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 72–82, March 2008.

[7] A. Gholami, "Sparse time-frequency decomposition and some applications," IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 6, pp. 3598–3604, June 2013.

[8] P. Flandrin and P. Borgnat, "Time-frequency energy distributions meet compressed sensing," IEEE Transactions on Signal Processing, vol. 58, no. 6, pp. 2974–2982, June 2010.

[9] L. Stankovic, S. Stankovic, I. Orovic, and M. G. Amin, "Robust time-frequency analysis based on the L-estimation and compressive sensing," IEEE Signal Processing Letters, vol. 20, no. 5, pp. 499–502, 2013.

[10] I. Volaric, V. Sucic, and S. Stankovic, "A data driven compressive sensing approach for time-frequency signal enhancement," Signal Processing, vol. 141, pp. 229 – 239, 2017.

[11] J. M. Bioucas-Dias and M. A. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," Image Processing, IEEE Transactions on, vol. 16, no. 12, pp. 2992–3004, 2007.

[12] J. Yang and Y. Zhang, "Alternating direction algorithms for $l1$-problems in compressive sensing," SIAM journal on scientific computing, vol. 33, no. 1, pp. 250–278, 2011.

[13] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," IEEE Access, vol. 3, pp. 490–530, 2015.

[14] V. Sucic, N. Saulig, and B. Boashash, "Estimating the number of components of a multicomponent nonstationary signal using the short-term time-frequency Rényi entropy," EURASIP Journal on Advances in Signal Processing, vol. 2011, no. 1, pp. 125–136, 2011.

[15] ——, "Analysis of local time-frequency entropy features for nonstationary signal components time supports detection," Digital Signal Processing, vol. 34, pp. 56 – 66, 2014.

[16] R. G. Baraniuk, P. Flandrin, A. J. E. M. Janssen and O. J. J. Michel. Measuring Time-Frequency Information Content Using the Rényi Entropies, IEEE Transactions on Information Theory, vol. 47, no. 4, 2001.

[17] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," IEEE Transactions on Information Theory, vol. 57, no. 7, pp. 4680–4688, July 2011.

[18] K. Qiu and A. Dogandzic, "Sparse signal reconstruction via ECME hard thresholding", IEEE Transactions on Signal Processing, vol. 60, no. 9, pp. 4551–4569, Sept 2012.

[19] I. Volaric and V. Sucic, "Localized Rényi Entropy Based Sparse TFD Reconstruction", Proceedings of the Second International Balkan Conference on Communications and Networking BalkanCom 2018 Podgorica, Montenegro, 2018. pp. 1-5

[20] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," Signal Processing, IEEE Transactions on, vol. 57, no. 7, pp. 2479–2493, 2009.

[21] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," Image Processing, IEEE Transactions on, vol. 19, no. 9, pp. 2345–2356, Sept 2010.

[22] S. Becker, J. Bobin, and E. J. Candés, "NESTA: a fast and accurate first order method for sparse recovery," SIAM Journal on Imaging Sciences, vol. 4, no. 1, pp. 1–39, 2011.

[23] L. Stankovic, "A measure of some time-frequency distributions concentration," Signal Processing, vol. 81, no. 3, pp. 621–631, 2001.

Maritime Communications and Remote Voyage Monitoring

Nobukazu Wakabayashi
Graduate School of Maritime Siences
Kobe University
Kobe, JAPAN
waka@kobe-u.ac.jp

Irena Jurdana
Faculty of Maritime Studies
University of Rijeka
Rijeka, Croatia
jurdana@pfri.hr

***Abstract*—For ships at sea, communicating with other ships or even to land-based sites has always been a very difficult task. Around 1900, the installation of wireless communication equipment on ships began. This form of communication initially used radio waves in the MF (Medium Frequency) band. Ever since it has been undergoing various technological innovations. At present, in addition to typical terrestrial communication in the VHF band, satellite communication using microwaves is also possible. Further, these forms of data communication have replaced the telegraph and telephone. Even now, however, the speed and capacity of data communication are significantly less than that of the land-based data communication network. After reviewing the situation of maritime communication, this paper studies the data communication for autonomous navigation of ships that is currently in demand focusing on remote monitoring. It appears that few instances of obtaining verification through detailed on-board data related to real-time communication have previously been possible. That being the case, the authors posit that it is worthwhile to demonstrate the feasibility of obtaining relevant verification based on actual data via a university training vessel equipped with the latest data collection system – a system that is relatively unavailable even on new, large-scale merchant ships. Particularly for digital data, maritime communication remains mainly via relatively slow and expensive narrowband satellite transmission. Though speedier and more efficient data transfer is possible through the use of simple data compression, the costs for this mode are currently prohibitively high. It is posited that economical, speedy, and efficient data transfer via data compression will increasingly become economically available for more ships in the future.**

Keywords— Maritime Communication, Data Communication, Digital Ship, Remote Monitoring

I. INTRODUCTION

It is common on land for information devices and information terminals to be connected to a network able to readily access the Internet. However, such access is not always possible for vessels offshore due to limiting satellite communication speed and cost factors.

From the past to the present, maritime communications have focused on matters related to distress and safety factors, for providing navigational support, and for communicating with the public. Symbolized by "SOS" calls over the years for assistance, distress and safety communications are those related to search and rescue operations for ships in distress. Communicating to provide navigational support refers to sharing information with nearby vessels as well as for exchanging information with port managers when entering or departing a port. The purpose of communicating with the public is to provide services similar to those on land using ordinary telegrams and telephones.

Communication is experiencing an increasing shift from analog to digital even at sea. The current requirement for data communication as part of maritime communication is also increasing. With the full implementation of the "Global Maritime Distress and Safety System (GMDSS)" in 1999 [1][2], a system that automatically sends digital distress signals via satellite has been used instead of telegraphing an "SOS" or sending a "Mayday" message by telephone. Digital radio communications such as AIS (Automatic Identification System) via VHF band terrestrial wave for exchanging navigational data between ships as well as between ships and coast stations were introduced [3], which is one of the modes of communication for navigational support. In public communication, data communication through maritime communication satellites have been realized as a service for the crew of cargo ships and passengers on luxury cruise ships, to send and receive e-mail, Web searching, and accessing SNS (Social Network Service), etc. via the Internet.

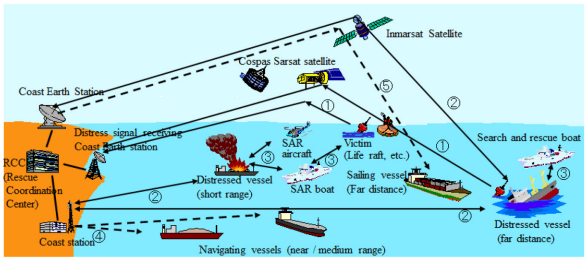
In this paper, the history and current situation of maritime communications – including difficulties – are described. Also, in preparation for the automatic navigation of ships, a technical study on the data communication required for the so-called "Digital Ship" is introduced [4]. It is hoped that through such input future maritime communications will be significantly improved.

II. HISTORY OF MARITIME COMMUNICATIONS

A. The beginning of maritime communications

A ship at sea can't connect directly to land-based sites or other ships via cable. It is therefore posited that wireless communication is the prime solution for effective communication. Before the development of wireless communications, ships at sea could only communicate within a visible distance and were restricted either to the use of various light forms and/or flags. Even today signal flags are an important means of ship-to-ship communication.

Ships began to be equipped with wireless communication devices at about the same time as Marconi's experiment on intercontinental wireless communication. At that time, steamships carrying passengers initiated an increasing demand for telegrams. However, there was no systematic arrangement for distress communications. As the "R.M.S. Titanic" was sinking in April 1912, a distress signal was transmitted by radio. And yet, the numerous casualties that



① Automatic transmission of distress signal (Satellite E-Pirb)
② Distress communication (DSC, MHF / VHF Transceiver, NBDP, Inmarsat, etc.)
③ Search and rescue communication (two-way radio telephone, VHF, etc.)
④ Safety information (NAVTEX)
⑤ Safety Information (Inmarsat EGC)

Fig. 1. Outline of GMDSS

resulted triggered the making of the treaty regarding the distress and safety communications for ships [5].

After that, the international VHF was primarily via voice communication. Since the implementation of GMDSS, it has been used for low-speed digital communication in the form of DSC (digital selective call).

B. Maritime distress and safety communications

The global safety at sea agreement conforms to the SOLAS Convention (Safety of Life at Sea) [6]. The International Convention on Maritime Search and Rescue takes into consideration search and rescue operations in the event of a maritime accident. The maritime communication system based on them is GMDSS (Global Maritime Distress and Safety System) [7] which:

- Realizes reliable reporting by automating and digitizing distress signals.
- Eliminates the need for special skills such as the transmission and reception of Morse code and makes effective use of simple buttons (Distress Buttons) and switches. This allows any member of the crew to respond in an emergency.
- Use satellite communications to secure stable communications.
- Stipulates maintenance of radio equipment to ensure reliable operation while at sea.
- Stipulates safety communication during normal operation.

Fig. 1 shows an overview of GMDSS in an image diagram.

C. Satellite communications

In former times and still in use today, long-distance communications at sea were limited to those using radio waves in the MF / HF band. However, in the latter half of the 20th century when communication satellites were activated, their use became widespread for maritime communication. Currently, mobile communication via “Inmarsat” satellite is available [8][9], as well as various other forms of communication. Initially, voice services were the main focus, but data communications are currently being enhanced. Communication via satellite is, however, still via narrow band with the result that the speed of effective communication is slower than that experienced on land. Besides, communication charges remain relatively expensive.

Regarding data communication via satellite, there was a time when narrow band digital data was exchanged by Inmarsat-B. Retired from service in December 2016, it had a low transmission speed of up to 128 kbps. Still in use today, however, is Inmarsat-C which is dedicated to such low-speed data communication as telex.

III. VDES – VHF Data Exchange System

Not only satellite communications, but also those in lower frequency bands increasingly use digital data modes rather than analog modes. The data communication method using wireless in the VHF band is called VDES (VHF Data Exchange System) [10]. VDES uses terrestrial waves and realizes only a relatively low communication speed. By 2007, ships of a certain size or more (Ex. Over 300G / T cargo ship in an international operation, etc.) were required to be equipped with AIS. AIS utilizes digital communication for transmitting voyage data collected by “own ship” in the 156 MHz radio band range. The data contained therein can be broadly divided into Dynamic Information, Static Information, and Voyage Related Information. The specific data items are as follows.

Position Report (Dynamic Information)

- MMSI (User ID)
- Navigation Status
- Rate of Turn
- SOG: Speed over Ground
- Position (Longitude, latitude)
- COG: Course over Ground
- Heading

Static and Voyage Related Information

- MMSI (User ID)
- IMO Number
- Call Sign
- Name of Ship
- Type of Ship & Cargo Type
- Overall Dimension and reference for position
- Estimated Time of Arrival
- Maximum Draught
- Destination

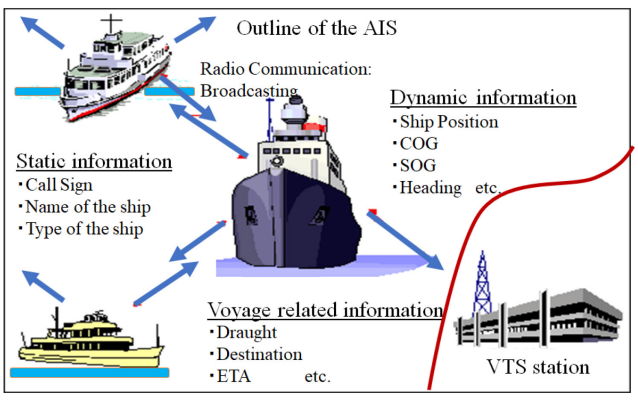


Fig. 2. Overview of using AIS

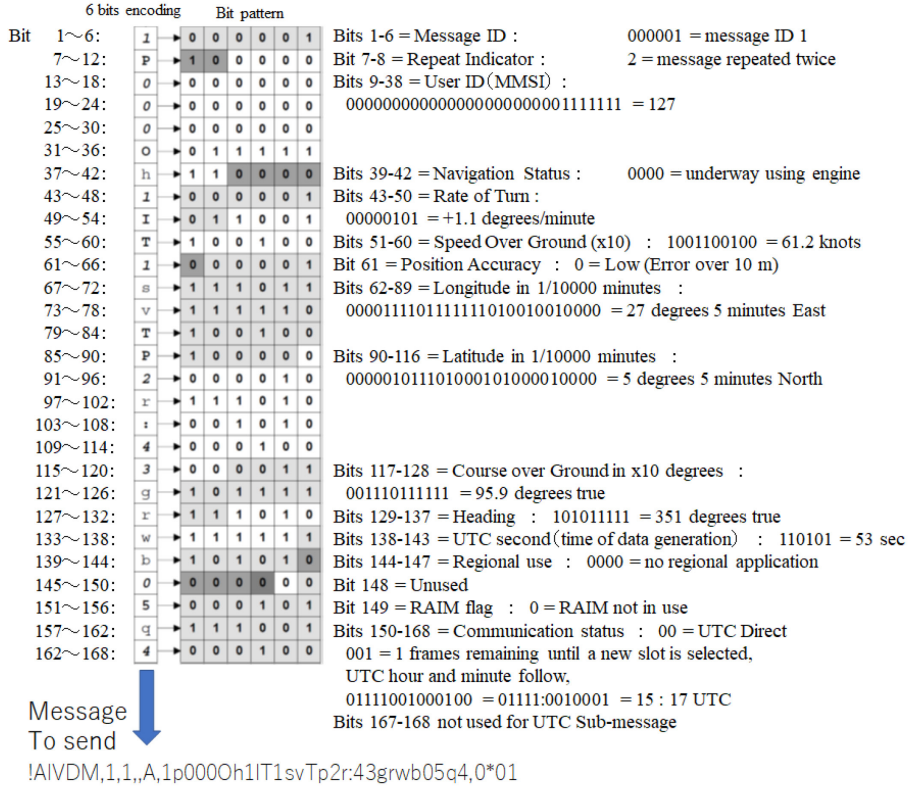


Fig. 3. AIS position report message of 1 record

During normal voyage conditions, a “Position Report” (Dynamic Information) is transmitted once every 10 seconds. However, the transmission cycle changes to reflect significant changes in voyage conditions. A “Static Information” and “Voyage Related Information” once every 6 minutes, are broadcast to surroundings. At the same time, AIS device can receive data from surrounding ships and identify the surrounding ships on its own ship. Fig. 2 shows overview image of AIS.

The transmitter power output of AIS is up to 12.5W while the signal can generally be received within a range of 30 to 40km in general. The encoding of AIS data transmission is slightly complicated so that the predetermined number of bits required to express numerical data is of variable length for each category. For example, 30 bits for MMSI, 4 bits for

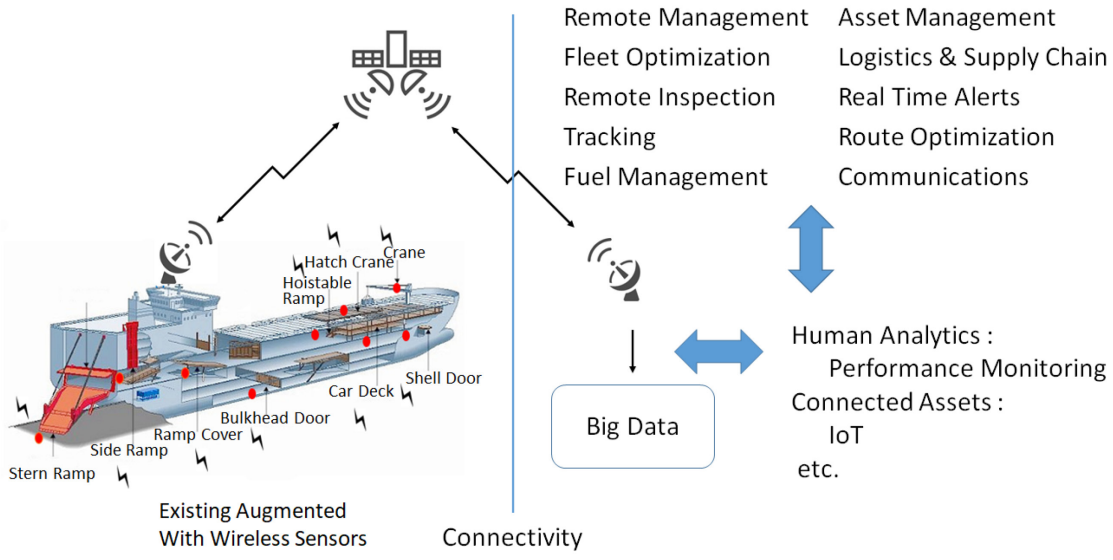


Fig. 4. Conceptual diagram of “Digital Ship”

Navigation Status, and 8 bits for Rate of Turn. All items are represented by integer values, and the items are sequentially converted totally to 168-bit binary data. For real number data, the number of digits after the decimal point (n digits) is determined in advance, and the original value is converted to an integer value by multiplying by 10 to the n th power. This is shown in Fig. 3. The bit pattern of this binary value is made to correspond to each character of the binary value every 6 bits, and the ASCII code corresponding to the character is finally transmitted. The encoding is performed every 6 bits to convert to actual alphanumeric characters and symbols, excluding the control characters of the ASCII code. Specifically, it converts decimal numbers 0-39 to hexadecimal numbers 0x30-0x57 and 40-63 in decimal to 0x60-0x77 (the prefix "0x" indicates a hexadecimal number). This is a slightly complicated conversion.

IV. Remote Voyage Monitoring

A. Digital Ship

Ongoing activities to further automate ship operation ships are also progressing. The authors have been studying the preparation for that [12][13][14]. Through the collection of a variety of voyage and shipboard factors, the movement towards more efficient marine transportation is progressing. Lloyd's Register calls such vessels "Digital Ships." Figure 4 shows the Digital Ship concept [15].

The AIS described above is for transmitting limited predetermined data to ships in the surrounding area as well as to coastal stations. However, a Digital Ship needs to transfer as much more data as possible. Also, since the data is expected to be transmitted to land-based operation management companies, satellite communications will be used instead of land-based equipment.

Previously, the authors independently developed TCS (Track Control System) [12] and have conducted experiments on actual ships. TCS can also be used for remote maneuvering – a function which is a primary step towards the further development of automatic navigation. Regarding the automation of ship operations, it is at the stage that concrete examples will be reported upon when they occur. Though no actual cases have yet been reported, concentrated upon so far have been factors related to collision avoidance methods.

B. Collection and transfering of voyage data

Being considered is concrete Voyage Data. A Motor Vessel, T.S. FUKAEMARU is a training ship operated by Kobe University that has been equipped with an advanced onboard Ethernet LAN – including Wireless LAN – for over 2 decades. All data regarding navigation, environmental factors, and the engine are collected in digital format via LAN [16][17][18]. The data types are as follows:

- Navigational

  GPS, GPS Compass, Gyro Compass, Magnetic Compass

- Environmental

  Anemometer, Current Profiler

- Engine related

  Temperatures, Pressure, Flow rate

- Targets

```
20190825120000006 5,$TIROT,-017.2,A*12
20190825120000009 2,$GPVTG,060.6,T,,,001.0,N,001.8,K*25
20190825120000021 3,$GPVTG,57.9,T,63.9,M,0.9,N,1.7,K*46
 …..

20190825120000368 5,$HEHDT,004.8,T*23
20190825120000370 2,$PFEC,GPhve,+0.048,A*18
20190825120000381 3,$PFEC,GPatt,5.1,+2.1,-2.1*4A
20190825120000401 5,$TIROT,-015.2,A*10
20190825120000411 0,$WIXDR,P,1.0111,B,0,C,26.6,C,0,H,69.
9,P,0,C,29.3,C,1*57
 …..

20190825120000980 2,$GPZDA,030000.00,25,08,2019,,*60
20190825120000981 1,$CCVTG,57.52,T,,,0.94,N,1.74,K*30
20190825120000981 1,$CCVLW,411.97,N,411.97,N*4d
20190825120000998 2,$GPGGA,025958,3215.1056,N,13309.9
152,E,2,05,01,+0006,M,+032,M,00,0000*6F201903201223459
81 3,$GPROT,0.0,A*31
```

Total 90 lines (= data records) in average for one second.

```
Explanation:
YYYYMMDDHHMMSSmmm D, NMEA Sentence
|  ||||||
|  |||||milli second
|  ||||  second
|  |||minute
|  ||hour
|  |day
|  month
Year

D: Device code of data source
```

Fig.5. An example of collected data (including timestamps and NMEA sentences) in certain one second

Radar TT, AIS

```
D: Device code of data source

 0 ... VDR (Voyage Data Recorder) or similar
       processor
 1 ... GPS (FURUNO)
 2 ... GPS Compass (JRC)
 3 ... GPS Compass (FURUNO)
 4 ... Doppler Sonner (DS-60)
 5 ... Gyro Compass (YOKOGAWA)
 6 ... Magnet Compass (NUNOTANI)
 7 ... EM Log (YOKOGAWA)
 8 ... Doppler Log (ATLAS)
 9 ... Accelerometer (MEMSIC)
 N ... Navigation data collecting equipment
 D ... Weather(Analog)
 A ... AIS
 R ... No.1 Radar (TT)
 S ... No.2 Radar (TT)
 W ... Weather Transmitter (VAISALA)
 E ... Engine Data Logger (TERASAKI)
 C ... CPU control system data
 H ... ADCP (Teledyne RD)
 B ... Sea Water Monitor
```

Fig. 6. Codes for devices to measure data source

The approximate numbers of these measurement data item (measurement channels) are as follows.

50 ch for navigational,	every 0.2 – 1 seconds
50 ch for environmental,	every 1 – 2 seconds
350 ch for engine related,	every 2 seconds

Approx. 450 Channels, total.

Fig. 5 shows an example of data collected for one second aboard T.S. FUKAEMARU. Each line shows one data record (one sentence sent from the device). From left to right are the Data Reception Time, the Data Measurement Device Identification, and the NMEA Sentence as received. In actual usage, the data with an average of 90 lines was received in 1 second and an abbreviated example is shown in Fig. 5. As described above, a large amount of Ship Operation Data is collected and received from all kinds of equipment. With the Data Collection System installed on the training ship, the amount of data during navigation is about 7.5 to 8 million lines (= data record) in one day (24 hours), and averages about 90 lines per second.

Fig. 6 shows a list of Measured Device Identifications in the recorded data. In this way, the system that collects and records the ship's operation data in real time onboard is unique to this training ship due to the fact that it is not generally used yet even for large commercial ships.

Note that various devices were used to collect such a large amount of data in just one second. When this is continued, the NMEA Sentence [19] received from each device will be transferred and recorded as text in 24 hours and the total volume will be approximately 500 MB. The simple calculation of the communication capabilities required to transfer this is as follows.

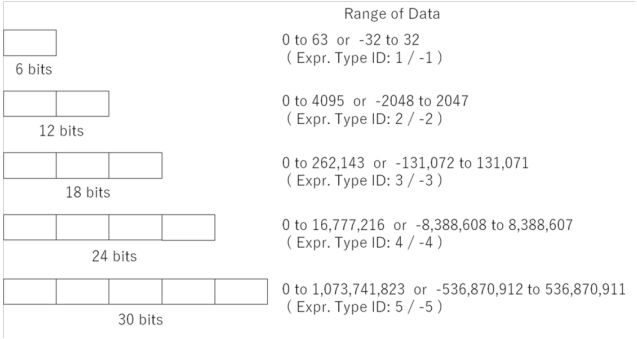


Fig. 7. Variable Length of six-bit word

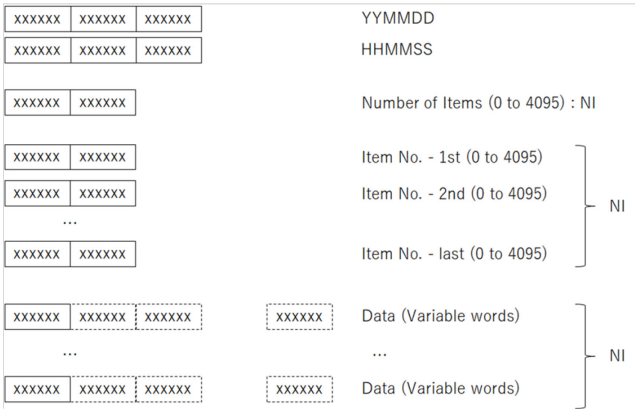


Fig. 8. Variable Length of six-bit word

Approx. 500 MB / day = 4,000 Mbit / day = 46 kbps

That is, a communication speed of 46 kbps is required. This is not impossible with current satellite communications. However, being considered is a reduction in the amount of data exchange in a more efficient format.

C. Examination of the data transfer format

Navigation data is numerical data. There are Binary and Text as manners of expressing numerical values in digital data on a computer. Even in the case of numerical data, ASCII text format is easier to understand, but in the case of multi-digit numerical values, Binary format reduces the number of bits required to express and is considered more efficient. However, a conversion procedure to binary is required – an almost negligible requirement in current computer performance.

When using binary data, it must be taken into account how many bits are used to represent the integer value. It is assumed that similar to the AIS transfer method, 6-bit encoding is finally performed and text is exchanged. It is necessary to consider the maximum expression range of each data item from 1 word to 5 words, with 6 bits as 1 word. The number of words used for each item is determined in advance. Fig. 7 shows the results.

If this is transmitted every time for all items, the amount of data for each transmission will be constant. However, the data measured by various devices on the ship is obtained at short intervals as described above. Some of them change almost every second, such as the latitude and longitude of the "own ship" position, heading, and rudder angle. On the other hand, there are some measurement items where the data does not change for several seconds to several tens of seconds. Not all items may need to be transferred every time. Therefore, as one method of reducing the amount of communication, consider transmitting only the items whose values have changed from the previous data. In this case, it is necessary to specify which data item to send each time. As a result, the data for one record will be formatted as shown in Fig. 8.

At that time, for each data item information such as the number of digits after the decimal point and how many words are used in the numerical expression are required. Since such input is constant, only this information was recorded separately according to each data item characteristic. "Data Item Table" is prepared in advance. Fig. 9 shows this table for T.S. FUKAEMARU. This table contains information about one data item per line. The data item number, the number of digits after the decimal point, the data item name, the unit, the number of words used to represent the data, and whether the data is positive only or positive or negative, and the assumed minimum and maximum values of the data are separated by commas in each line.

The bit pattern of the resulting binary data is converted to text for every 6 bits corresponding to a part of the ASCII code and is used as one record data.

This conversion adopted a slightly different method from the encoding used in AIS. In the ASCII code shown in Fig. 10, 1-word data is 6 bits, so 64 characters are required. Except for control characters and other characters used in sentences, "! (Exclamation point)," ", (Comma)," and "* (Asterisk)" are assigned to the colored area in Fig. 10. As a result, it can be

Item No., digits under DP, Name, Unit, Expr.Type, Min., Max.

20,0," GPS Time(hhmmss)","",3,000000,235959,
21,0," GPS Time(msec)","",2,000,999,
22,0,"GPS Status","",2,65,86,
23,1,"GPS Time Difference","hours",-4,-12.5,12.5,
24,0,"GPS(UTC)
Date[YYYYMMDD]","",4,20000101,20301231,
25,2,"GPS(UTC)
Time[hhmmss.xx]","",4,000000.00,235959.99,
26,0,"GPS Latitude(deg)","deg",-2,-90,90,
27,6,"GPS Latitude(min)","min",5,00.000000,59.999999,
28,0,"GPS Longitude(deg)","deg",-2,-180,180,
29,6,"GPS
Longitude(min)","min",5,00.000000,59.999999,
30,1,"GPS COG","deg",2,000.0,360.0,
31,1,"GPS SOG","kts",-2,-99.9,99.9,
32,1,"GPS Mag","deg",2,000.0,360.0,
33,1,"GPS Variation","deg",-2,-90.0,90.0,
34,0,"GPS Altitude","m",-3,-99,9999,
35,0,"GPS No. of Satellite","",1,0,50,
36,1,"GPS HDOP","",2,0.0,20.0,
…
…

Fig. 9. Data item table

assigned to 64 consecutive characters and can be converted by simply adding 0x30 to a 6-bit binary value. Processing is simpler than AIS coding.

Finally, the checksum is calculated by performing the exclusive OR operation on the ASCII code one character at a time from the character following "!" to the last character.

Fields in one line are a timestamp (HHMMSSsss) consisting of hours, minutes, seconds, and milliseconds, a serial number when transmitting data divided into multiple lines, a character string encoded in 6-bit units, and "*" followed by a 2-digit hexadecimal checksum. Except for checksum, these fields are separated by a comma. If the encoded character string is long so that the entire line is 80 characters or less, it is divided and transmitted every 60 characters. This is shown in Fig. 11. Fig. 12 shows an example of this processing.

ASCII Code Table

Lower Digit \ Upper Digit	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
a	LF	SUB	*	:	J	Z	j	z
b	VT	ESC	+	;	K	[	k	{
c	FF	FS	,	<	L	\	l	|
d	CR	GS	-	=	M	]	m	}
e	SO	RS	.	>	N	^	n	~
f	SI	US	/	?	O	_	o	DEL

Fig. 10. Range for 6bit encoding in ASCII code table

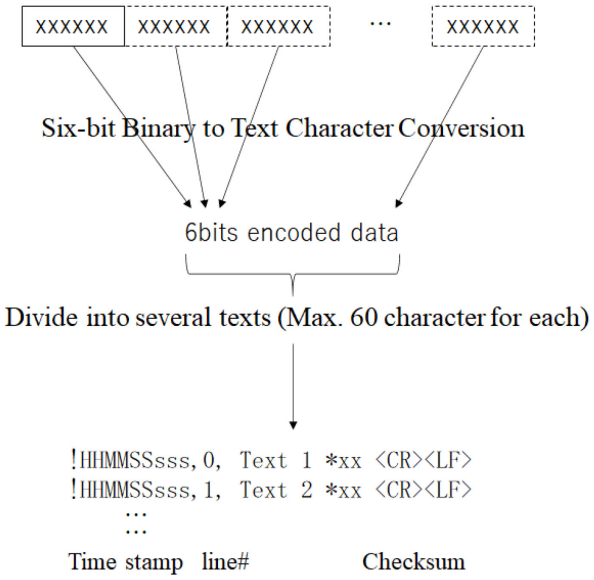


Fig. 11. Range for 6bit encoding in ASCII code table

In the method of transmitting only the difference (Send only the data that has changed since the last transmission), if there is a possibility that some items with no data may occur if reception starts in the middle, all items are sent even if there is no change at regular intervals. In the next experiment, all item's data are sent every one minute.

D. Conversion Efficiency Experiment using recorded data

To evaluate the efficiency of the proposed conversion method, the conversion efficiency was measured using recorded data. During the 2019 voyage of the T.S. FUKAEMARU, data was collected 24-hours a day for the 22 days. Fig. 13 shows the results.

Fig. 13 shows each case: 1) fixed-length CSV format, 2) fixed-length Binary format, 3) change items only, data items and values (text format), 4) only the change items as binary values and 6-bit coding proposed above. It shows the required transmission rate (bps) as a percentage of the average data transfer rate when the original NMEA sentence is sent as it is. In the method of sending only the change data, all items are sent once a minute.

In the NMEA sentence, about 35 kbps (= 35000 bps) is required, whereas in the proposed format, it is about 1.2 kbps (= 1200 bps), the data transfer rate is reduced to about 3.4% on average, and the standard deviation is 159 bps (0.4%) for 22 days data.

E. Estimating data traffic for general use

The above is a discussion regarding the amount of remote monitoring data for Digital Ship. Following is an estimate of the data amount of communication classified as general public communication. Considered is a standard oceangoing cargo ship with about 25 crew members and a luxury cruise ship with a total of 3,000 crew members.

For cargo ships,

- Work-related communication is the main focus
- For vessels operating in new forms, voyage data and control data are focused upon

Original NMEA Sentences

20190825120000006 5,$TIROT,-017.2,A*12
20190825120000009 2,$GPVTG,060.6,T,,,001.0,N,001.8,K*25
20190825120000021 3,$GPVTG,57.9,T,63.9,M,0.9,N,1.7,K*46
20190825120000025 2,$GPRMC,025957,A,3215.1056,N,13309.9150,E,001.0,061.,250819,,*22
20190825120000035 2,$GPDTM,W84,,0.0,N,0.0,E,+0.0,W84*44
20190825120000052 2,$GPGSA,A,3,17,19,06,28,23,00,00,00,00,00,00,00,04.2,01.4,03.8*09
20190825120000052 E,$000008,39035860,282,20028,4221,21017,1449,436,424*31
20190825120000071 2,$GPGSV,3,1,10,17,70,349,54,19,54,330,50,06,41,296,50,28,41,203,47*7E
20190825120000081 1,$GPGGA,030000,3215.1086,N,13309.9192,E,1,11,1.5,21,M,,M,,*77
20190825120000081 1,$CCVTG,61.23,T,,,0.97,N,1.80,K*3b
20190825120000081 1,$CCVLW,411.96,N,411.96,N*4d
20190825120000089 2,$GPGSV,3,2,10,03,40,043,42,23,31,100,46,09,25,144,41,22,19,044,44*7D
20190825120000096 3,$GPZDA,030000,09,01,2000,-09,00*65
20190825120000108 2,$GPGSV,3,3,10,01,13,069,42,11,12,095,39,00,00,000,00,00,00,000,00*77
20190825120000120 2,$GPGLL,3215.1056,N,13309.9150,E,025957.00,A*0A
20190825120000120 2,$CCVTG,65.84,T,,,0.97,N,1.79,K*34
20190825120000120 2,$CCVLW,412.24,N,412.24,N*4d
20190825120000148 2,$GPHDT,004.3,T*32

… 80 lines in total for 1 second for example.

Differential Binary (Variable Length for one record) six-bit encoded

!120000006,0,^IYMC006500D0E0F0G0H0I0J0K0L0M0N0O0P0R0S0T0U0V0W0X0Y0Z0[0\0]*13
!120000006,1,0^0_0`0a0b0c0d0f0g0h0i0j0k0l0m0n0o101@1A1B1C1E1F1G1J1K1L1M1N*5f
!120000006,2,1O1T1U1V1X1Y1Z1[1\1]1^1_1`1a1b1c1d1e1f1g2P2Q2R2S2T2U2V2W2X2Y*19
!120000006,3,2Z2[2\2]2^2_2`2a2b2c2d2e2i2k2n2o303133344\4]4^4_4a4b4c4d4e4f*02
!120000006,4,4g4h4i4j4k4l4m4n4o505152535455565758595:5;5<5=5>5?5@5A5B5C>4*59
!120000006,5,>5>6>7>9>:>>?X?YA=A>A?A@AAABACADAEAFAGAHAIAJAKALAMANAOAPAQB9*0f
!120000006,6,B:B;B<B=BaBbBcBdBeBfBgBhBiBjBnBoC0C1C2C3C4C5C6C7C8C9C:C;C<C=*0b
!120000006,7,C>C?C@CACBCCCDCECFCGDDDEDFDGDHDIDJDNDODPDQDRDSDTDUEkElEmEnEo*20
!120000006,8,F0F1F2F3F4F5F6F7F8F9F:F;F<F=F>F?F@FEFFFGFHFIFJFSFTFUFVFWFXFY*1d
!120000006,9,FZF[F\F]F^F_F`FaFbFcFdFeFfFgFhGMGNGOGPGQGRGSGTGUGVGWGXGYGZG[*24
!120000006,10,G\G]G^G_GaGbGcGgGkGmGnGoH5ITI\JUJVJXJYJZJ[J\J]J^J_*76

... Just 11 lines for one second (for same time as above NMEA sentences).

Fig. 12. An example of data processing by means of proposed method

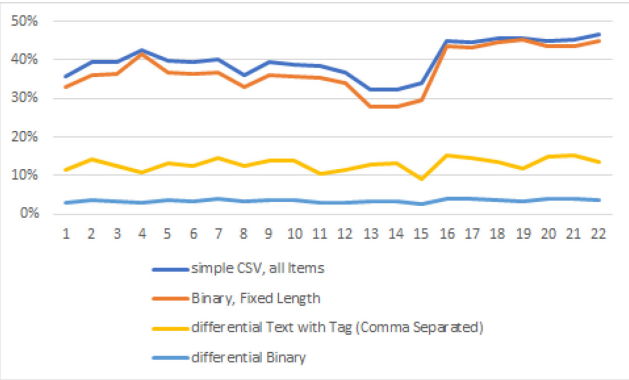


Fig. 13. Efficiency of data conversion in 4 cases

In addition to the above, it is assumed that crew members will personally have access to the Internet. On luxury cruise ships, it is assumed that many passengers, as well as the crew, have demands for Internet access equivalent to that available on land.

An example of the estimate (per person) for light user:

Movie: 150MB

SNS: 50MB

Mail: 30MB

Web browsing: 20MB

per person per day – estimated to be approximately 250MB in total. As a result, a rough estimation of communication volume is

250 MB・3000 person = 750 GB;
Bandwidth ≈ 70 Mbps (100% use)

250 MB・750 person = 187.5 GB;
Bandwidth ≈ 17.4 Mbps (25 % use)

125 MB・25 person = 3.125 GB;
Bandwidth ≈ 289 kbps (crew only, 50% use)

in rough estimation respectively.

It may be possible at present for a cargo ship crew of 20 to 30 persons per ship to experience a network environment like that available on land. More than 3000 passengers and crew members aboard luxury cruise ships, however, are

unavoidably restricted from accessing the Internet due to narrowband satellite communications. Thus, it can be seen that on a ship, usage is far from equivalent to a land-based network environment. Immediate improvement of the situation is desired.

V. Conclusions

This paper offers an condensed overview of the historical development of maritime communications. So far it seems that there have been few cases where verification utilizing detailed data obtained from various sources onboard ships related to real-time communication for remote voyage monitoring is possible. The authors believe that it is worthwhile to show the feasibility of such a development based on actual data obtained on a university training vessel equipped with the latest data collection system. Currently, such a Data Collection Function is seldom to be found even in new, large-scale merchant ships. At present, maritime communications – especially those of digital data – mainly use the relatively slow and expensive narrowband satellite communication method. In today's world, however, it has been found that data transfer is possible simply and inexpensively via data compression. Due to the fact that data communication charges are currently relatively high, it is hoped that more ships will incorporate the data compression method.

The "Digital Ship" concept proposed by Lloyd's Register was briefly described. As one of the factors, a ship to/from land communication protocol for remote monitoring was examined. Efficiencies were calculated using recorded data from actual ship results. That being the case, the simulation showed that the required communication speed could be reduced to about 3% or 4% of the original format of sentences.

Looking to the future, the authors would like to verify the effectiveness of actual communication between a ship and the land over some time.

Acknowledgment

The authors wish to express their gratitude to Paul H. Faust, Jr., Professor Emeritus of Tezukayama University in Nara, Japan for suggestions regarding English usage.

References

[1] IMO, "Shipping Emergencies - Search and Rescue and the GMDSS," Focus on IMO, 1999.

[2] IMO, "HARMONIZATION OF GMDSS REQUIREMENTS FOR RADIO INSTALLATIONS ON BOARD SOLAS SHIPS," COMSAR/Circ.32, 2004.

[3] ITU-R, "Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band," Recommendation ITU-R M.1371-5, 2014.

[4] IMO, "Autonomous shipping," http://www.imo.org/en/MediaCentre/HotTopics/Pages/Autonomous-shipping.aspx (Accessed Mar. 2020).

[5] IMO, "History of SOLAS (The International Convention for the Safety of Life at Sea)", http://www.imo.org/en/KnowledgeCentre/ReferencesAndArchives/HistoryofSOLAS/Pages/default.aspx (Accessed Mar. 2020).

[6] IMO, "International Convention for the Safety of Life at Sea (SOLAS), 1974," http://www.imo.org/en/About/Conventions/ListOfConventions/Pages/International-Convention-for-the-Safety-of-Life-at-Sea-(SOLAS),-1974.aspx (Accessed Mar.2020).

[7] IMO, "Radiocommunications," SOLAS Chapter IV, 1974.

[8] Inmarsat, "Maritime, Inmarsat official site," https://www.inmarsat.com/maritime/ (Accessed Mar. 2020).

[9] Inmarsat, "Services, Inmarsata official site",https://www.inmarsat.com/services/ (Accessed Mar. 2020).

[10] Francisco Lazaro Blasco, Ronald Raulefs, Wei Wang, Federico Clazzer, and Simon Plass, "VHF Data Exchange System (VDES): An enabling technology for maritime communications," CEAS Space Journal, Vol. 11, No. 4,2017.

[11] IMO, "AIS transponders," http://www.imo.org/en/OurWork/Safety/Navigation/Pages/AIS.aspx (Accedded Mar. 2020).

[12] T. Watanabe, N. Wakabayashi, M. Urakami, and D. Terada, "Development of Track Control System utilizing Heading Control System for Ocean Observation Sailing," Proc the Twenty-seventh (2017) International Ocean and Polar Engineering Conference, San Francisco, ISOPE, 529-536.

[13] N. Wakabayashi, T. Watanabe, M. Urakami, and D. Terada, "Development of Simple Dynamic Positioning System —Algorithm and User Interface—," Proc the Twenty-seventh (2017) International Ocean and Polar Engineering Conference, San Francisco, ISOPE, 507-512.

[14] N. Wakabayashi, T. Watanabe, M. Urakami, and Y. Yano, "Tablet Control System for Offshore Support and Research Vessel — Development, Implementation, and Operational Testing—," Proc the Twenty-eighth (2018) International Ocean and Polar Engineering Conference, Sapporo, ISOPE, 922-928.

[15] Lloyd's Register, "Digital Ships - Procedure for assignment of digital descriptive notes for autonomous and remote access ships," ShipRight Design and Construction, 2019.

[16] N. Wakabayashi, Y. Yano, S. Shiotani, and K. Murai, "Sailing Data Transfer and Display System Using Wireless LAN in the Bridge and its Application to Navigation Support System," 11th IAIN (International Association of Institutes of Navigation) World Congress, CD-ROM, October 2003.

[17] N. Wakabayashi, T. Watanabe, M. Urakami, and Y. Yano, "Vessel LAN - Design, Implementation, and Operation," 2018 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications, Graz, Austria, CobCom, July 2018.

[18] M. Fujii, M. Hayashi, M. Urakami, and N. Wakabayashi, "The Development of Meteorological and Oceanographic Data Collection and Recording System Operating on Training Ship," ASME 2014 33rd international Conference on Ocean, Offshore and Arctic Engineering, San Francisco, OMAE2014-23883, June 2014.

[19] National Marine Electronics Association , "NMEA 0183 Standard for Interfacing Marine Electronic Devices Version 3.01," 2002.

Evaluation of data sets for mobile radio signal coverage up to 150 meters above ground

1st Klaus Kainrath
Institute of Aviation
FH JOANNEUM GmbH
Graz, Austria
klaus.kainrath@fh-joanneum.at

2nd Jakob Feiner
Institute of Internet Technologies & Applications
FH JOANNEUM GmbH
Kapfenberg, Austria
jakob.feiner@fh-joanneum.at

3rd Wilhelm Zugaj
Institute of Internet Technologies & Applications
FH JOANNEUM GmbH
Kapfenberg, Austria
wilhelm.zugaj@fh-joanneum.at

4th Erich Leitgeb
Institute of Microwave and Photonic Engineering
Graz University of Technology
Graz, Austria
erich.leitgeb@tugraz.at

5th Holger Fluehr
Institute of Aviation
FH JOANNEUM GmbH
Graz, Austria
holger.fluehr@fh-joanneum.at

6th Mario Gruber
Institute of Aviation
FH JOANNEUM GmbH
Graz, Austria
mario.gruber@fh-joanneum.at

***Abstract*—For the integration of unmanned aircraft into civil airspace, Europe-wide harmonized regulations have been adopted which will come into force in the EU member states starting in July 2020. There are currently no specific design regulations for Unmanned Aerial Vehicles (UAVs). This also applies to the data link that is to be used initially for control and subsequently for payload, such as video feeds. In order to be able to use mobile radio technology as a data link for UAVs, it must first be investigated whether it is suitable for this purpose. To accomplish this, measurements were carried out in the air using UAVs to evaluate the signal quality at operating heights of up to 150 meters above ground. The results are discussed in this paper.**

Keywords—UAV, BVLOS, mobile radio, data analysis

I. Introduction

Unmanned aircraft are developing into a rapidly evolving aviation sector with high potential for various economic sectors in the European Union. The aviation act amendment 2014 regulated UAVs in Austrian law, which led to an approval according to LBTH 67 [1] at Austro Control for operation according to §24 of the aviation act. The European Commission adopted EU-wide regulations on technical requirements for UAVs in 2019 [2]. These regulations are divided into three categories (open, specific and certified) with different safety requirements appropriate to the risk (Fig. 1). EU-wide regulations will thus also come into force in Austria in July 2020. At the technological level, however, the EU-wide regulations do not provide additional rules, such as specific frequency bands for communications, flight control architectures or collision avoidance.

II. UAV Datalink

A. Official Publications

The International Telecommunication Union (ITU) report ITU-R M.2171 [3] deals with spectrum requirements for the communication of unmanned aerial systems (UAS), Air Traffic Control (ATC)-UAS, Command and Control (C2) and Sense & Avoid. It is stated that terrestrial and satellite-based systems will probably be required for the operation of UAS. Based on this, ITU-R M.2233 [4] deals with the technical characteristics regarding Beyond Visual Line of Sight (BVLOS) applications for Medium/Large UAS (M/LUAS) and Small UAS (SUAS). Draft plans for frequency usage for UAVs are known from the European Conference on Postal and Telecommunications Administration (CEPT) Workshop on Spectrum for Drones/UAS in 2018 [5]. However, concrete frequency bands for BVLOS applications are not specified or foreseeable.

Fig. 1. EASA regulations DR 2019/945, Source: Adapted from [6]

The Radio Technical Commission for Aeronautics (RTCA), with its Special Committees 228 (SC-228), develops the Minimum Operational Performance Standards (MOPS) for Unmanned Aerial Systems (UAS) [7]. The available publications include recommendations for hardware for Command and Control (C2) and data links. Bands for C2 link use are also covered including L-band (1-2 GHz), C-band (4-8GHz) and Satcom in multiple bands. A distinction is made between Phase 1 MOPS in L- and C-band for Visual Line Of Sight (VLOS) operation and Phase 2 MOPS Satcom for BVLOS operation. The Joint Authorities for Rulemaking of Unmanned Systems (JARUS) also published the Remotely Piloted Aircraft System (RPAS) Required C2 Performance concept [8], which, in accordance with International Civil Aviation Organization (ICAO) document 10019 RPAS Manual [9], describes conceptual considerations for the highest level of abstraction of a C2 link without going into technological aspects in detail.

However, no hardware specifications have been officially recommended to date. Although MOPS exist, UAV operators cannot access commercially available hardware that can be easily approved by Austro Control in order to get an operation certificate.

B. Problem Statement

Since there is currently no standardized hardware for a UAV data link, model flight components on the one hand and proprietary hardware and mobile radio on the other hand are currently used. For VLOS operation of UAVs standard remote controls are sufficient. Due to the introduction of new regulations, however, flights BVLOS, which are currently legally not possible in Austria, will also become an issue in the near future. The problem with proprietary hardware for BVLOS operations is that there is no infrastructure, and due to the power limitation in the spectrum, it is not possible to transmit at any transmission power. Thus, the mobile network offers a good technological basis for UAVs.

C. Mobile Radio Technology

The 3rd Generation Partnership Project (3GPP) is responsible for providing a stable environment of specifications for wireless technologies. It brings together seven telecommunications standardization organizations worldwide. Mobile standards are provided as release documents that are constantly being revised. There are five major generations in the development of mobile communication technologies. Each generation uses different carrier frequencies and usually also different technologies for signal transmission and access procedures. This results in different requirements for the network structure for each generation. There are also differences within a generation in different countries. The frequencies available in Austria are presented by Rundfunk und Telekom Regulierungs-GmbH (RTR) [10]. All generations have the same basic components: these include the mobile device, e.g. a mobile phone, the base station (BS) and the backbone network. The largest part of the network is called backhaul. It covers all connections of all base stations to the core network. These connections are usually implemented with fiber optic cables, but in rural areas microwave links are preferred. In mobile communications technology, the wireless data transmission is between Mobile Equipment (ME) such as mobile phones and Base Transceiver Stations (BTS) [11]. Due to the limited range of a radio link, the mobile network is divided into radio cells. In each of these radio cells there is a base station that is connected to the network backbone. The form of a wireless network is designed as a repetitive structure and built as a honeycomb of cells, with the base station in the middle. This structure allows the repeatedly use of available frequencies, as identical frequencies can be reused if the cells using the same carrier frequency are spaced apart and do not overlap.

D. Is mobile radio suitable?

Currently, mobile radio technologies up to the fourth generation, Long Term Evolution (LTE), are designed exclusively for terrestrial use. Considering the radiation characteristics of a base transceiver station (BTS), this limitation results from the shape of the main beam and its power density (Fig. 2). However, several side and reflected beams (called side lobes) also have an upward radiation pattern.

Mobile radio signals do not experience any reflections in the air, only free space losses, atmospheric influences and, to a lesser extent, scattering. This leads to the assumption that the coverage of a mobile radio cell at altitudes up to 150 meters above ground level is still sufficient to have a proper data connection for UAVs.

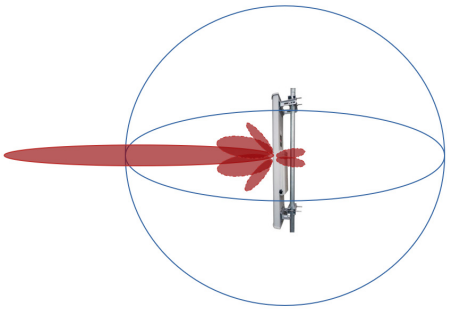


Fig. 2. Radiation pattern LTE antenna

State-of-the-art mobile radio networks are optimized exclusively for terrestrial users. The interference caused by other neighboring BTS is lower than the attenuation of the signals caused by infrastructure like buildings. However, these attenuations do not exist for users located high above the transmitting stations (UAVs) where often a VLOS connection is possible. It is unclear to what extent the interference (e.g. reflected beams) in the airspace is low enough as well as the signal quality is good enough to ensure secure communication for UAVs.

III. Data Aquisition

A. Long Term Evolution Parameters

The second generation [2G] mobile radio network is not suitable for UAVs due to its low bandwidth and the third generation (3G) will be switched off in the near future (prob. 2021 [12]). For this reason, the measurements focus primarily on the fourth generation (4G) mobile radio network.

With 4G the OFDM Modulation scheme is used. There are Cell Reference Signals (CRS) used that are inserted at regular intervals within the OFDM time frequency grid (Fig.3). The four resource elements per resource block that are dedicated to the reference signal, as shown in Fig. 3.

The position of the CRS in the resource block is unique for each cell and is encoded in the physical cell ID. The physical cell ID is recognized at the login process at the base station by the User Equipment (UE). The Primary Synchronization Signal (PSS) and Secondary Synchronization Signal (SSS) are used in this process. In addition, the transmission of the cell reference signals is carried out with increased Transmission (TX) power [11]. The UE measures the current reception status as the Reference Signal Received Power (RSRP). The RSRP then is calculated by:

$$RSRP = \frac{1}{n}\sum_{k=1}^{n} P_{rs,k} \quad (1)$$

This value is the average power (in Watt) received from a single Reference Signal (RS) resource element.

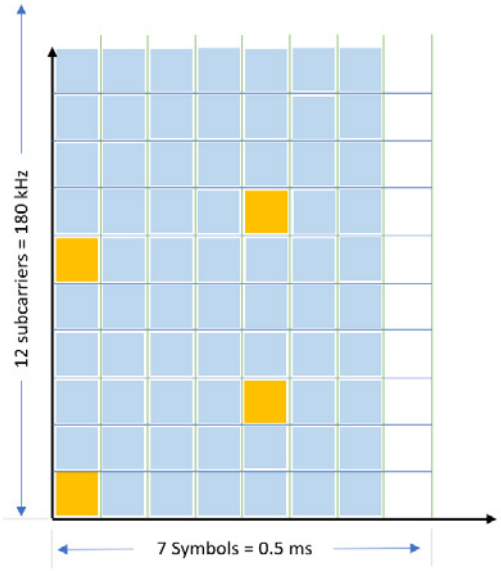


Fig. 3. LTE OFDM-Slot with highlited Reference Blocks (RB) [11]

RSRP only measures the RS power and excludes all noise and interference power. By knowing the absolute RSRP value from the cell, the UE can calculate the downlink path-loss and even estimate the distance to the base station. The UEs send acknowledge signals back to the cell, so the base station can dynamically adjust the power. The maximum RSRP value is based on the maximum input power of a UE and converted to dBm equals to -44dBm. The minimum value is -140dBm.

RSRP serves as an important parameter for quantifying the reception strength. The cell-specific reference signals according to 3GPP "TS 36.211" [13] are used for this purpose. If the UE, e.g. mobile phone, or LTE dongle, can receive reference signals, a precise measurement of the RSRP value is possible. Unfortunately, the RSRP value does not give an information about the signal quality. So, there is still no indication of the strength of the reference signal compared to the overall energy in the channel, which is also known as the Receive Signal Strength Indicator (RSSI).

Therefore, the Reference Signal Receive Quality (RSRQ) value is introduced. The RSRQ value determines if the actual signal quality is sufficient, or if there are too many interferences and there should be a handover. For determine the RSRQ value, the power of one whole OFDM symbol (see figure 3) is measured. The value received is the Received Signal Strength Indicator (RSSI). The RSRQ value is then calculated by the following equation:

$$RSRQ = \frac{RSRP}{(\frac{RSSI}{N_{RB}})} \quad (2)$$

N_{RB} is the number of Resource Blocks and the RSSI parameter represents all the received power including the wanted power from the serving cell as well as all co-channel power and other sources of noise. So, measuring the RSRQ becomes important near the edge of the cell, when it comes to the decision whether a handover should be performed or not, regardless of absolute RSRP. The maximum value of RSRQ is -3dB and the minimum value of RSRQ is -20dB.

B. Data Aquisition

In order to verify the assumption that the mobile radio signal is still adequate for the use as C2 data link for UAVs at altitudes up to a maximum of 150 meters Above Ground Level (AGL), measurement flights were carried out using commercial UAVs.

Classical high-frequency measuring instruments cannot be transported. Also, the option to carry a Software Defined Radio for the measurement in the UAV is not possible, because these require a lot of computing power, a certain supply voltage and it needs special configuration for measurement and logging. Therefore, the measurements were performed with smartphones. The test flights were carried out alternately with one and two devices. The equipment used consists of the models listed in Table I:

TABLE I. USER EQUIPMENT

Model	Main UE specification	
	Frequency (MHz)	*LTE bands*
Xiaomi A2 Lite	800, 850, 900, 1700, 1800, 2100, 2300, 2600	1, 2, 3, 4, 5, 7, 8, 20, 38, 40
Xiaomi Redmi 5 Plus	800, 850, 900, 1700, 1800, 2100, 2300, 2600	1, 3, 4, 5, 7, 8, 20, 38, 40

With the help of the Android application G-Net Track Pro [14] relevant data of the mobile network can be recorded. The Application Programming Interface (API) has full access to the UEs modem and can log many parameters such as RSRP, RSRQ, RSSI etc. not only of the serving cell but also from the neighbouring cells.

IV. MEASUREMENT DATA ANALYSIS

A. Preparation of the Data

The acquisition of the data was carried out in the recent 2 years. The collected data is available as text files, containing parameters, such as time, position, RSRP, RSRQ, RSSI, the serving and neighbouring cells and many more.

The Programming language Python was used to evaluate this data. To prepare the data for analysis, a file is created for each flight in the evaluation software. Flight data must be in a format similar to CSV, where the data is separated by a separator. The first line contains the titles of the respective columns. Every other line represents a recording point of the measurement flight and contains parameters like time stamp, the current geo-coordinates (longitude, latitude), transmission power parameters (RSRP, RSRQ, Signal to Noise Ratio (SNR)), neighbouring cells properties etc. [14].

The first step is data cleansing. For example, rows with missing values are deleted and column types are adjusted. Data points that deviates too much from the expected values is filtered out. After the clean-up, the next step is to add additional columns with data needed for the plot creation. For example, a column with the flight altitude AGL is calculated to display the flight path. The transmission power parameters obtained by the APP (see [14]) are each divided

into intervals and stored in new columns accordingly (RSRP, RSRQ, SNR).

To support the interpretation of the plots, the intervals of the measured values RSRP, RSRQ and RSSI were separated into 6-level scales [17]. These are shown in the Table II:

TABLE II. MEASUREMENT VALUE SCALING

RSRP Level	**RSRP scaling**	
	Value in dBm	***Description***
1	-50 to -65	Excellent
2	-65 to -80	Very good
3	-80 to -95	Good
4	-95 to -110	Fair
5	-110 to -125	Bad
6	-125 to -140	Very bad
RSRQ Level	**RSRQ scaling**	
	Value in dB	***Description***
1	-3	No interference
2	-4 to -5	Minimal interference
3	-6 to -8	Little interference
4	-9 to -11	Moderate interference
5	-12 to -15	Strong interference
6	-16 to -20	Extreme interference
RSSI Level	**RSSI scaling (approx. taken from [16])**	
	Value in dBm	***Description***
1	-25 to -49	Maximum possible data rate can be attained
2	-50 to -72	Maximum data rates can be achieved
3	-73 to -79	Normal data rate possible
4	-80 to -89	Reduced data rate
5	-90 to -104	Interruptions possible
6	-105 to -117	Almost no stable connection possible

B. *Plot generation*

The data sets prepared can then be used to create plots. Especially, values of interest like RSRP, RSRQ are 2-dimensional (2D) plotted against time, or altitude. Furthermore, 3-dimensional (3D) plots were generated, using imported maps from OpenStreetMap (OSM). Within these 3D plots, one can see the path of the measurement flight plotted as sequence of points representing coded values of the desired parameter.

V. RESULTS

A. *Static Measurements*

The static measurements are intended to show the extent to which the measured values of the smartphones in use differ. For this purpose, terrestrial measurements were carried out. In these tests, measurements were made with identical and different smartphone models (Table I). In several runs the position of the UEs to each other was also varied. The measurements were carried out in urban areas. Therefore, the mean of the RSRP values are significantly lower than the RSRP values near a single rural base transceiver station. The recorded results were plotted over time as seen in Fig. 4.

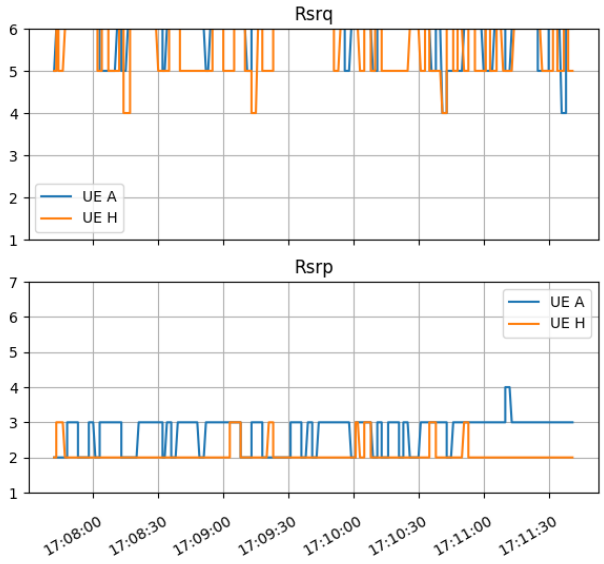


Fig. 4. Measurement results of two identical UEs in the same position

For a better overview, the measurement parameters for RSRP and RSRQ are summarized in the table below:

TABLE III. STATIC MEASUREMENT RESULT

Parameter	**2 identical UEs in the same position**	
	UE1 (scaling)	***UE2 (scaling)***
Average RSRP	-93.18 (3)	-90.22 (3)
Deviation RSRP	1.77	1.96
Correlation factor	-0.11	
Average RSRQ	-17.04 (6)	-15.53 (6)
Deviation RSRQ	2.20	1.94
Correlation factor	0.24	

The RSRQ value of this measurement is very bad. This is due to the fact that these measurements were carried out on the ground in urban area, the base transceiver station was not in the line of sight and it was almost 1km away from the UE. Comparison tests with only one UE showed on average similar values. In addition, the two mobile phones that were located next to each other were registered in the same mobile network at the same BTS which has caused additional interferences. Since this was measurable for both UEs, the standard deviation can be taken from this measurement as a reference for the dynamic measurements.

B. *Dynamic Measurements*

For the dynamic measurements, measurement flights were performed with a UAV as described in Chapter III B. Rural base transceiver stations were selected for the evaluation. The obtained results show less interference than in urban areas. Furthermore, from a legal point of view it would not be allowed to fly with the UAV over inhabited areas. The following plots (Fig. 5 to 8) compare the results

for two providers at a base station in Lower Austria.

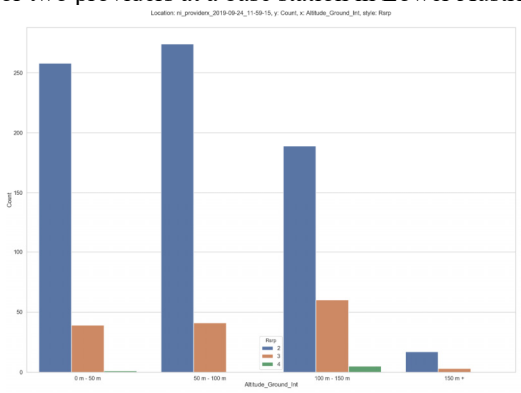


Fig. 5. RSRP values in relation to altitude – Provider X

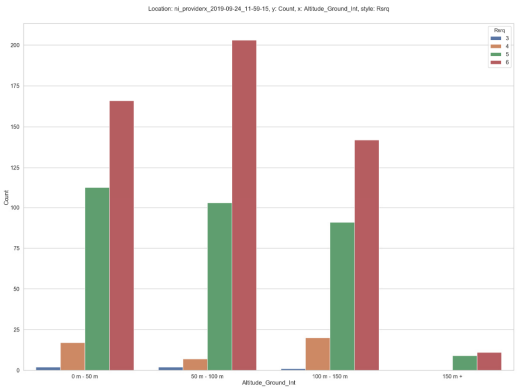


Fig. 6. RSRQ values in relation to altitude – Provider X

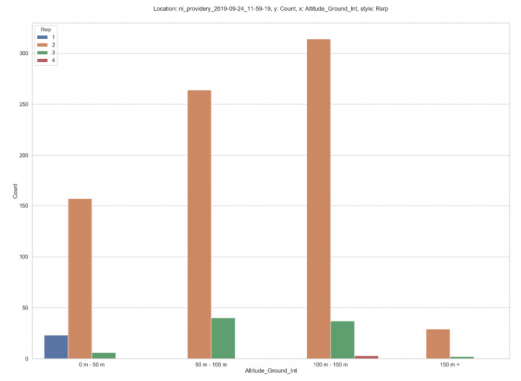


Fig. 7. RSRP values in relation to altitude – Provider Y

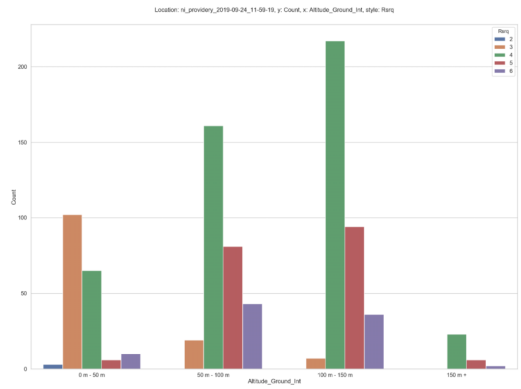


Fig. 8. RSRQ values in relation to altitude – Provider Y

Finally, a 3D plot is shown in Fig. 9. This shows the course of the RSRP value over the time of a provider. You can follow the flight and see the altitude profiles.

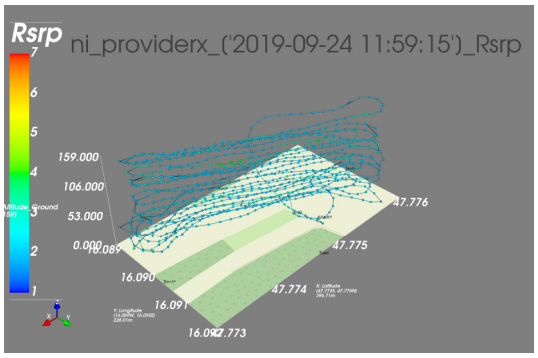


Fig. 9. 3D Plot of the RSRP values in relation to altitude – Provider X

VI. Conclusions and Outlook

A. Conclusion

With the evaluation of the first plots, it quickly becomes clear that the signal coverage is sufficiently available up to 150m above ground level (AGL). Static measurements show a slight variance of the measured values, according to the model and position. This depends, among other things, on the installed antenna, on the shape and material of the housing and, accordingly, on the position of the UE relative to the base station. To take these deviations into account, you have to take a look at the exact measured values and not only at the scaling. For example, the RSRP shows a standard deviation from the average value of around 2dBm with peaks up to 10dBm.

The plots generated from the dynamic measurements show significant differences in the reception quality of the two investigated providers at the measurement location. Provider X shows a high RSRP value over the entire altitude, but the corresponding quality (RSRQ) is rather poor. This indicates a very noisy signal in this particular LTE band. This is different for Provider Y. There, the average RSRP value is lower, but the RSRQ value shows good signal quality throughout. The RSSI value measured for both providers indicates good overall signal (Level 2, Table II) strength in the applied LTE band.

B. Outlook

The results so far show that the pure signal coverage is provided in heights up to 150m AGL. Continued investigations must now be carried out, which also examine the bandwidths. It must be evaluated how the network utilization influences the signal quality. Data transfer tests are planned for this purpose. Furthermore, the same measurements will be carried out on different days to evaluate the variation of the network. Moreover, it is recommended to carry only one mobile phone for the dynamic measurements in the UAV in the future, since they may influence each other.

Based on the further static measurements performed, an approximation for standard deviation of the measured values is should be generated, which can be considered in further evaluation. In the course of this work it is also planned to generate a mathematical model of the signal quality by means of a software like Matlab [15]. This would provide a

theoretical model. Thus, the measured values can be compared to the simulated values and evaluated.

REFERENCES

[1] Austro Control, LBTH67, Betrieb von unbemannten Luftfahrzeugen, https://www.austrocontrol.at/jart/prj3/austro_control/data/uploads/LFA/LTH_LFA_ACE_067.pdf, online, last seen 13.04.2020

[2] EASA - Civil drones (Unmanned aircraft), https://www.easa.europa.eu/easa-and-you/civil-drones-rpas, last seen 13.04.2020

[3] Characteristics of unmanned aircraft systems and spectrum requirements to support their safe operation in non-segregated airspace, https://www.itu.int/en/ITU-R/space/snl/Documents/R-REP-M.2171-2009-PDF-E.pdf, last seen 13.04.2020

[4] Examples of technical characteristics for unmanned aircraft control and non-payload communications links, https://www.itu.int/en/ITU-R/space/snl/Documents/R-REP-M.2233-2011-PDF-E.pdf, last seen 13.04.2020

[5] CEPT Workshop on Spectrum for Drones / UAS, https://cept.org/ecc/tools-and-services/cept-workshop-on-spectrum-for-drones-uas, last seen 13.04.2020

[6] New EASA Drone Regulations 2020, DR 2019/945 | IR 2019/947 https://www.grupooneair.com/new-easa-drone-regulations/, last seen 13.04.2020

[7] RTCA Special Committee 228 Minimum Performance Standards for Unmanned Aircraft Systems, https://www.rtca.org/sites/default/files/sc-228_sept_2019_tor.pdf, last seen 13.04.2020

[8] RPAS "Required C2 Performance" (RLP) concept, http://jarus-rpas.org/sites/jarus-rpas.org/files/storage/Library-Documents/jar_doc_13_rpl_concept_upgraded.pdf, last seen 13.04.2020

[9] Manual on Remotely Piloted Aircraft Systems (RPAS) (Doc 10019), http://store.icao.int/products/manual-on-remotely-piloted-aircraft-systems-rpas-doc-10019, , last seen 13.04.2020

[10] Austrian Regulatory Authority for Broadcasting and Telecommunications, https://www.rtr.at/de/tk/Frequenzen, online, last seen 13.04.2020

[11] M. Sauter, Grundkurs Mobile Kommunikationssysteme, 6th ed, Springer, 2015, p231-273.

[12] News report, Der Standard, https://www.derstandard.at/story/2000107163399/die-tage-der-3g-handys-sind-gezaehlt, last seen 13.04.2020

[13] 3rd Generation Partnership Project;Technical Specification Group Radio Access Network;Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 13), https://www.arib.or.jp/english/html/overview/doc/STD-T104v4_10/5_Appendix/Rel13/36/36211-d20.pdf, last seen 13.04.2020

[14] G-Net Track Pro, Gyokov Solutions, https://www.gyokovsolutions.com/manuals/gnettrackpro_manual.php, last seen 13.04.2020

[15] Matlab, https://www.mathworks.com/products/matlab.html, last seen 13.04.2020

[16] LTE-Anbieter-Info, ASU - Arbitrary Strength Unit, https://www.lte-anbieter.info/technik/asu.php, last seen 13.04.2020

[17] [16] LTE-Anbieter-Info, RSRP (Referenz Signal Received Power), https://www.lte-anbieter.info/technik/rsrp.php, last seen 13.04.2020

Searching for the Optimal Design of Small Payment Accessories

Mladen Pesic[1,2], Stephan Rampetzreiter[1], Walther Pachler[1], Holger Arthaber[2]
[1]*Development Center Graz, Infineon Technologies Austria AG*
[2]*Institute of Electrodynamics, Microwave and Circuit Engineering, TU Wien*

***Abstract*—Small payment accessories (e.g., watches, bracelets, rings) are becoming the future of "contactless" payment. Designing these devices as passive (batteryless) in terms of power supply is faced with challenges concerning miniaturization and compliance to standards. To evaluate performance, we introduce a simulation framework that can predict a design's minimum operating magnetic field strength (H_{min}) with an accuracy under 0.1 A/m. The framework combines S-parameter models of the device's antenna and the ISO-standardized setup (the ISO test PCD assembly) with a data-based nonlinear model of the device's IC. Techniques for optimizing the energy transfer are discussed (tuning and power matching) and backed by analytical and practical examples. We also demonstrate how to use the simulation framework to determine the impact of the device's structure on energy transfer. Two designs of small payment accessories are ultimately compared, both as models in the framework and as fabricated samples. By applying power matching instead of tuning, the second design's size can be reduced by approximately half, without significant change in H_{min}. As the predicted H_{min} values match the measurements, the results show that multiple design parameters can be varied within the framework to determine their effect on H_{min}, which is of great assistance for finding the optimal design.**

I. Introduction

Today, the evolving trend of digitalization has also left its mark on the payment process. While cash payments will probably not be rendered obsolete anytime soon, it is the "contactless" payment method that is going through strong transformation. Beside the established bank card and especially the mobile phone in the last decade (via NFC), other payment devices have emerged, such as wearable small *payment accessories*: watches, bracelets, rings, keychains, etc [1]. Common for contactless payment methods is their basis on proximity coupling RFID systems operating at 13.56 MHz [2].

The driving force behind small payment accessories is "tokenization", which has gained momentum since the emergence of cryptocurrency [3]. Possibility to utilize a small wearable item for payment instead of cash or bank cards has redefined the term "wallet", which is already in use for cryptocurrency containers [4]. Tokenization also demands miniaturization, which puts emphasis on optimizing the energy transfer, especially for passive devices. Even though active devices (NFC) and antenna "boosting" techniques are also directed at payment applications [5]–[8], this work focuses on modeling and optimization of their passive counterparts.

Compliance to standards is of great importance for proximity coupling systems, with EMVCo being the relevant standard for payment [9]. As dimensions go down, certain standards' requirements become more challenging for the energy transfer. In passive small payment accessories design, special focus must be given to compliance. Even when it is fulfilled, manufacturers strive for optimal performance so they can gain advantage over competitors. Since small form factor antennas are not defined as strictly as the ones in bank cards [10], [11], there is some "uncharted territory" as to which design is optimal. Also important is to establish a meaningful benchmark for the design's performance and decide if there is need for a redesign. Typical benchmark for contactless payment is the minimum operating magnetic field strength (H_{min}), given the basis on inductive coupling and magnetic field strength (H) being the quantity that describes it [2].

If the designer decides there is room for improvement and a redesign is needed, it usually means a geometry parameter (e.g., antenna size, number of windings, copper track width and gap) or a material property (e.g., relative permittivity, loss tangent) have to be changed. Multiple redesigns are often needed, creating a loop with resource-consuming iterations, where reducing their number would be advantageous. This could be possible with a simulation framework able to approximate actual test setups for performance benchmarks such as H_{min}. In this work, we elaborate on modeling and design concepts for proximity coupling systems utilized as small payment accessories. We also demonstrate how our simulation framework can be used to predict H_{min}. By varying multiple design parameters, this framework can also be used to "search" for the optimal design of small payment accessories.

II. Simulation framework

Proximity Integrated Circuit Card (PICC) is an established term for proximity coupling transponders operating at 13.56 MHz [11]. Although small payment accessories are transponders not usually in card form, here we still refer to them as PICCs for historical reasons. Analogously, *Proximity Coupling Device* (PCD) is the term used to describe the reader.

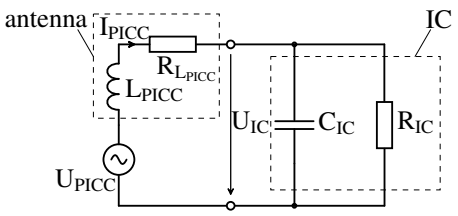


Fig. 1: Simplified equivalent circuit of a PICC

In its simplest form, a PICC circuit model consists of an antenna, represented by inductance L_{PICC} and its serial resistance $R_{L_{PICC}}$, and an IC, represented by a parallel connection of capacitance C_{IC} and resistance R_{IC} (Fig. 1). The "source" voltage U_{PICC} is induced by the PCD.

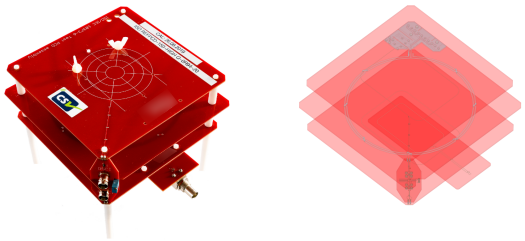


Fig. 2: The ISO test PCD assembly: apparatus and 3D model used for simulation

The *ISO test PCD assembly* (Fig. 2) is a standardized setup developed by ISO for performing various tests in PICC verification [12]. Among other applications, the ISO test PCD assembly allows observing a certain voltage (so-called *calibration coil* voltage U_{CC}) that is directly proportional to H and can therefore be used to measure H_{min}. In order to model this H_{min} measurement, we created a simulation framework that contains three components: the ISO test PCD assembly, the antenna design, and the IC. The first two are represented by linear S-parameter models, whereas for the nonlinear IC we used an empirical approach and developed a data-based model [13], [14]. Combination of all three, with addition of a source and an impedance matching network, constitutes the entire simulation framework (Fig. 3).

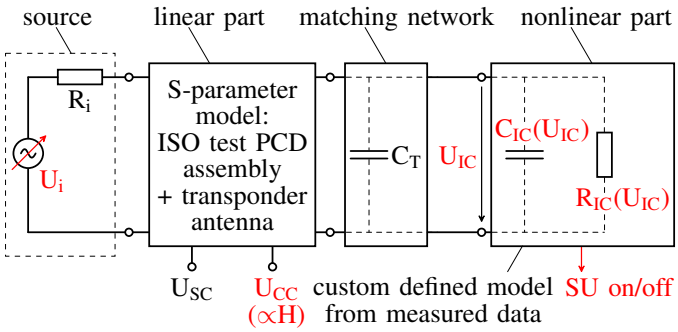


Fig. 3: Circuit model of the simulation framework

For the linear part, an FEM-based 3D model was created in Ansys HFSS for the ISO test PCD assembly and the transponder's antenna design as a DUT placed on top of it. A library of various designs was created, which can be quickly interchanged as DUTs. An example with a small payment accessory as DUT is depicted in Fig. 4. Field simulation of the model results in an S-parameter block, which can be further used in circuit simulations.

The highly nonlinear IC can be modeled based on its input characteristics as voltage-dependent capacitor $C_{IC}(U_{IC})$ and resistor $R_{IC}(U_{IC})$ connected in parallel. Using an LCR meter, these input characteristics can be measured and the results can be used to develop a custom defined nonlinear circuit component (Fig. 3) that describes the dependence of C_{IC} and R_{IC} on U_{IC}. In addition, a test feature of ICs in

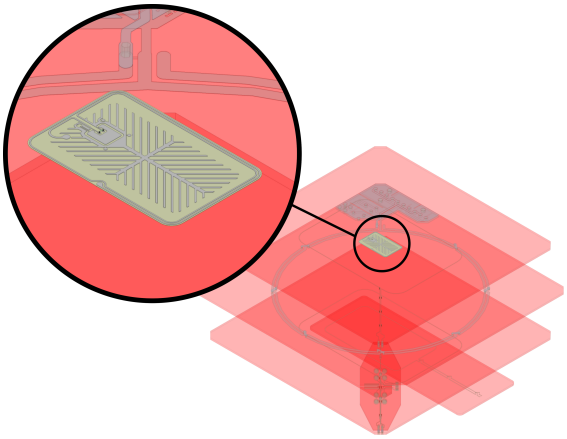


Fig. 4: Model of a small payment accessory as a DUT on the ISO test PCD assembly

production allows a digital signal to be read out from one of the IC's pins. This so-called *start-up* signal (SU) toggles in the exact moment when the IC is turned on. It can also be acquired during the LCR meter measurement and expressed as a function of U_{IC}, making it visible within the framework.

Sweeping the source voltage U_i (Fig. 3) and converting U_{CC} to H using the known proportionality factor allows H of the entire transponder to be *observable*. Thus, SU can be expressed in relation to H, where the H value at which SU changes from 0 (SU off) to 1 (SU on) is designated as H_{min}. The true H_{min} actually lies in the range between SU off and SU on, but we decided to use the latter for the H_{min} definition (Fig. 5). The framework is therefore able to predict H_{min} for any potential antenna design combined with an IC with known input characteristics. It can be especially useful for reducing the number of fabricated prototypes within a redesign loop. Accuracy of the H_{min} prediction is evaluated by actual H_{min} measurements and is confirmed to be below 0.1 A/m.

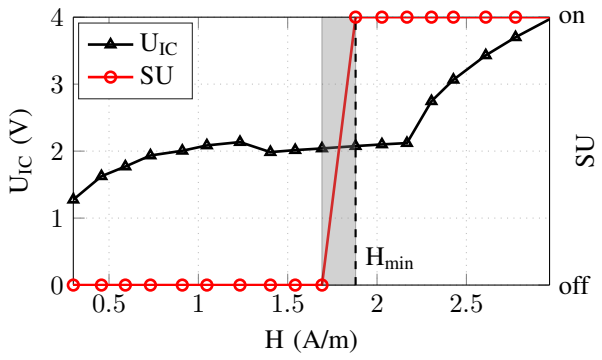


Fig. 5: Small payment accessory modeled with the simulation framework: U_{IC} and SU over H (with H_{min} clearly visible)

III. Optimizing Energy Transfer

During design, antenna and IC need to be combined in such a way that the PICC can draw as much power as possible from the PCD. This is particularly important for small form factor antennas because the lower coupling coefficients between PCD and PICC must be compensated [15]. As the principle of inductive coupling relies on resonant circuits, resonance frequency (f_{RES}) is an important factor [16], [17]. The most common definition of f_{RES} is when the PICC fulfills the *phase resonance* criterion

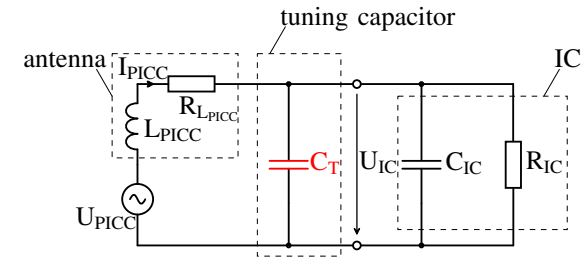
$$\Im\mathrm{m}\{\underline{Z}_{PICC}\} \stackrel{!}{=} 0\,, \tag{1}$$

where f_{RES} has a form similar to the Thomson equation

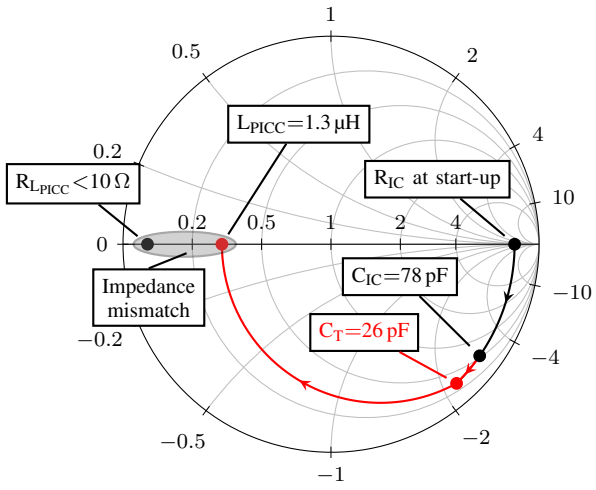
$$f_{RES} = \frac{1}{2\pi\sqrt{L_{PICC}C_{IC}}}\,. \tag{2}$$

The simplest way for a PICC (Fig. 1) to achieve phase resonance is by placing a parallel tuning capacitor C_T between the antenna and the IC (Fig. 6a). Solving (1) for C_T gives

$$C_T = \frac{R_{IC} - 2\omega^2 C_{IC} R_{IC} L_{PICC} + \sqrt{R_{IC}^2 - 4\omega^2 L_{PICC}^2}}{2\omega^2 R_{IC} L_{PICC}}\,. \tag{3}$$



(a) Equivalent circuit



(b) Smith chart plot

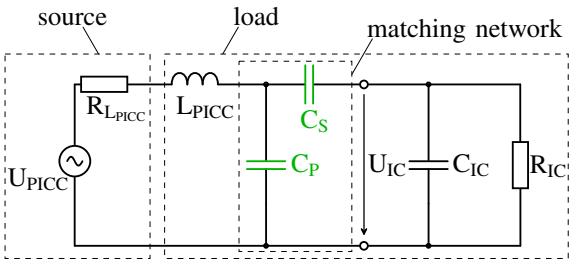
Fig. 6: PICC with a parallel tuning capacitor

For a case with known circuit component values (e.g., L_{PICC}=1.3 µH, C_{IC}=78 pF, and typical R_{IC} value at the IC's start-up), evaluating (3) results in C_T=26 pF. The tuning process is also depicted in a Smith chart with impedance trajectories for 13.56 MHz (Fig. 6b). Viewing the PICC as an RF network, the "source" impedance is actually $R_{L_{PICC}}$, which is usually under 10 Ω for wire-embedded copper antennas. Phase resonance does not necessarily mean that optimal energy transfer is reached, because a mismatch remains between $R_{L_{PICC}}$ and the load's (consisting of Z_{IC} in parallel with C_T) resistive part, which would need to be equal if optimal energy transfer is targeted. This mismatch can be mitigated using power matching techniques.

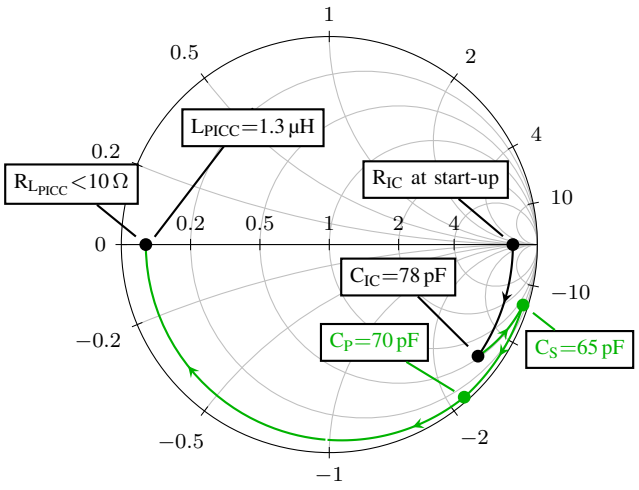
For typical power matching of an RF system, a matching network is placed between the source and the load. When it comes to PICCs, $R_{L_{PICC}}$ assumes the role of the source impedance. Since L_{PICC} is inseparable from $R_{L_{PICC}}$, the matching network needs to be inserted between the antenna and the IC, not between source and load, strictly speaking. Therefore, the matching network itself becomes part of the load (Fig. 8a). It usually consists of two capacitors: serial (C_S) and parallel (C_P). Because antennas are more easily adaptable in terms of design, power matching is performed from the IC side, so that either C_P or C_S can be connected first to the IC, resulting in two different constellations: "parallel first" and "serial first". For both cases, the condition

$$\Re\mathrm{e}\{\underline{Z}_{load}\} \stackrel{!}{=} R_{L_{PICC}} \tag{4}$$

must be met in addition to phase resonance.



(a) Equivalent circuit



(b) Smith chart plot

Fig. 7: PICC with a power matching network: "serial first"

Solving (4) for the case of "serial first" (Fig. 7a) leads to analytical expressions

$$C_S = \frac{\omega C_{IC} R_{IC}^2 R_{L_{PICC}} + \sqrt{R_{IC} R_{L_{PICC}} (\omega^4 C_{IC}^2 R_{IC}^2 L_{PICC}^2 + \omega^2 C_{IC}^2 R_{IC}^2 R_{L_{PICC}}^2 + \omega^2 L_{PICC}^2 + R_{L_{PICC}}^2 - R_{IC} R_{L_{PICC}})}}{\omega R_{IC} (\omega^2 L_{PICC}^2 + R_{L_{PICC}}^2 - R_{IC} R_{L_{PICC}})} \quad (5)$$

$$C_P = \frac{\omega R_{IC} L_{PICC} - \sqrt{R_{IC} R_{L_{PICC}} (\omega^4 C_{IC}^2 R_{IC}^2 L_{PICC}^2 + \omega^2 C_{IC}^2 R_{IC}^2 R_{L_{PICC}}^2 + \omega^2 L_{PICC}^2 + R_{L_{PICC}}^2 - R_{IC} R_{L_{PICC}})}}{\omega R_{IC} (\omega^2 L_{PICC}^2 + R_{L_{PICC}}^2)} \quad (6)$$

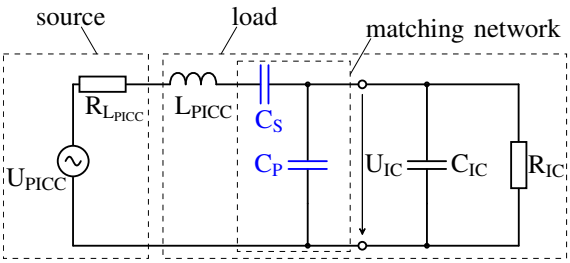
for C_S and C_P.

In the Smith chart (Fig. 7b), the "serial first" procedure for the example with known values starts from C_{IC}=78 pF with connecting a serial capacitor with C_S=65 pF, followed by a parallel capacitor with C_P=70 pF. From there, connecting the antenna with L_{PICC}=1.3 µH exactly arrives at the desired source impedance $R_{L_{PICC}}$.

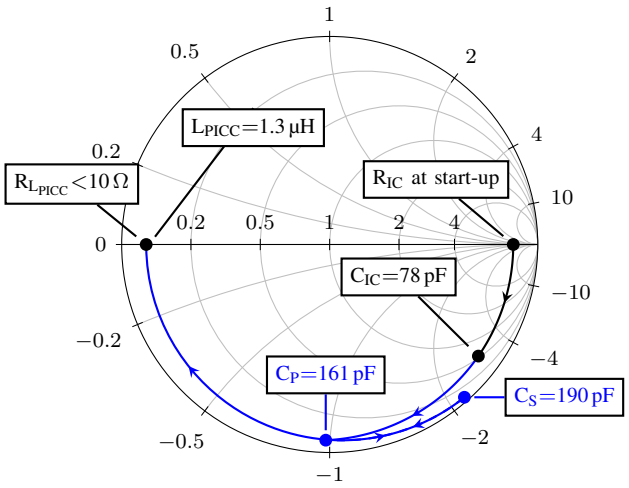
On the other hand, solving (4) for the "parallel first" case (Fig. 8a) gives

$$C_S = \frac{\omega L_{PICC} + \sqrt{R_{L_{PICC}} (R_{IC} - R_{L_{PICC}})}}{\omega^3 L_{PICC}^2 - \omega R_{L_{PICC}} (R_{IC} - R_{L_{PICC}})} \quad (7)$$

$$C_P = \frac{\sqrt{R_{L_{PICC}} (R_{IC} - R_{L_{PICC}})} - \omega C_{IC} R_{IC} R_{L_{PICC}}}{\omega R_{IC} R_{L_{PICC}}}. \quad (8)$$



(a) Equivalent circuit



(b) Smith chart plot

Fig. 8: PICC with a power matching network: "parallel first"

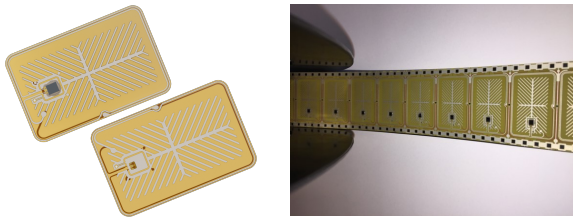
In the Smith chart (Fig. 8b) the "parallel first" procedure starts with a parallel capacitor with C_P=161 pF connected to the IC, followed by a serial capacitor with C_S=190 pF in order to arrive at the same point as the "serial first" case, from where the antenna is connected in the same manner.

Whether to use the "parallel first" or "serial first" approach depends on the application. Obviously, the "parallel first" case requires significantly higher capacitance values, meaning that the "serial first" approach saves area on the PICC by using physically smaller external capacitors. On the other hand, high values of these capacitors make the matching procedure less dependent on C_{IC} variation due to production tolerances.

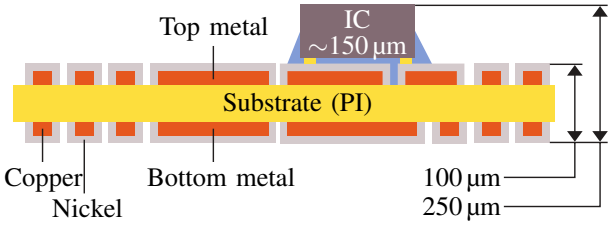
It should be noted that despite f_{RES} also having other definitions [18], the described approaches remain the same, only the resonance criterion (1) needs to be changed accordingly.

IV. Design Example

Infineon's SPA (Smart Payment Accessory) module (Fig. 9) is used as an example for discussing the various techniques in designing small payment accessories [19]. It is based on a patented "fishbone" parallel plate structure used as a built-in tuning capacitor (Fig. 9a) [20]. Starting from a very thin polyimide (PI) substrate, copper is deposited on both sides to fabricate the antenna coil and the matching network as a parallel plate capacitor with PI as the dielectric (Fig. 9b). This two-level structure allows more antenna windings than its one-sided counterpart would for the same device area. A thin layer of nickel is deposited over copper as a passivation layer and the IC is assembled into the module using flip chip technology.



(a) Top and bottom side (left) and tape reel (right)



(b) Cross section

Fig. 9: Infineon's SPA module

The motivation behind the built-in capacitor is to avoid external matching components and manufacture a PICC where the energy transfer is already optimized. The fabrication costs are thereby reduced, as well as fabrication times, because assembly of additional surface-mounted components is not needed. However, the reason for the "fishbone" design instead of a simple rectangular plate capacitor is an example of how the layout's structure and material properties play an important role in optimizing the energy transfer. Sometimes, thinking only in terms of circuit modeling and matching capacitance values may not be enough, and one needs to rely on more accurate modeling techniques, such as the aforementioned simulation framework.

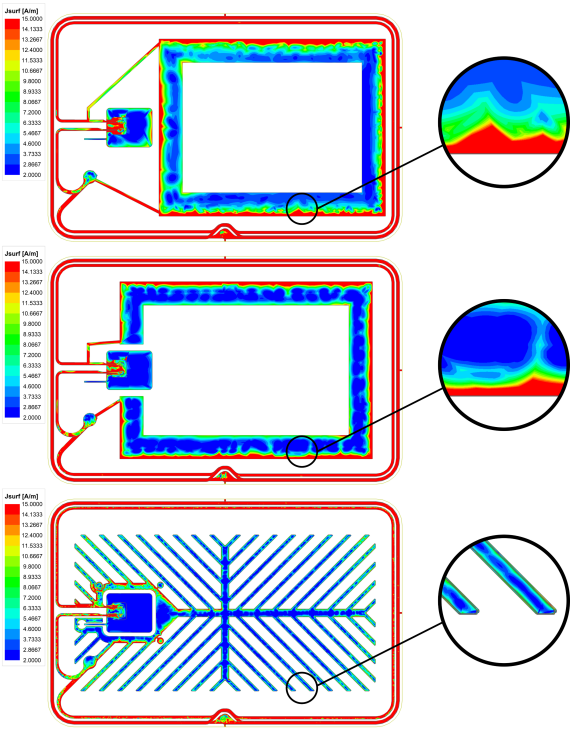


Fig. 10: Surface current comparison: enclosed frame-shaped design (top), frame-shaped design with slit (middle), and "fishbone" design (bottom)

The quality factor (Q) of the antenna, which is essentially an inductor [21], is determined as

$$Q = \frac{\omega L_{\mathrm{PICC}}}{R_{\mathrm{L_{PICC}}}}. \tag{9}$$

Since both the antenna and the capacitor are made of copper, the magnetic flux responsible for inducing $\mathrm{U_{PICC}}$ also passes through the capacitor and induces eddy currents in the capacitor's plate that, as per Lenz's law, tend to oppose the magnetic flux responsible for their origin [22]. These eddy currents result in more power dissipation, thus increasing $\mathrm{R_{L_{PICC}}}$ in (9) and ultimately reducing Q. If the matching capacitor were made as a full plate concentrated in the middle of the module, it would significantly reduce the mechanical stability and make the thin module less robust against multiple bends that occur in daily use. Even if the capacitor were made as a rectangle framed-shaped plate and spread out over the majority of the module's area, it would still suffer from higher losses caused by eddy currents. Adding a slit in the frame would prevent the eddy currents from enclosing themselves along the frame's shape and would force them to travel a longer path. Introducing more such slits into the parallel plate structure eventually leads to the "fishbone" design, where eddy currents need to travel a much longer distance in order to enclose themselves. The simulation framework was used to illustrate this phenomenon (Fig. 10). The area of the "fishbone" design was estimated and two further designs were modeled, one as an enclosed rectangular frame-shaped plate and another with the same shape that also includes a slit. Both of these were intenionally chosen to have poorer performance, in order to emphasize the role of layout's geometry for such designs. For the same purpose, the capacitor area was kept constant for all three designs. After field simulation, eddy currents can be visualized in a surface current plot (Fig. 10). It can be seen that, as expected, eddy currents are the strongest in the enclosed frame-shaped design, whereas they are the weakest for the "fishbone" design.

Fig. 11 shows the magnetic field profiles of the three designs. The effect of Lenz's law is especially noticeable in the enclosed frame-shaped design, where the magnetic field is significantly reduced by eddy currents in the area surrounded by the antenna.

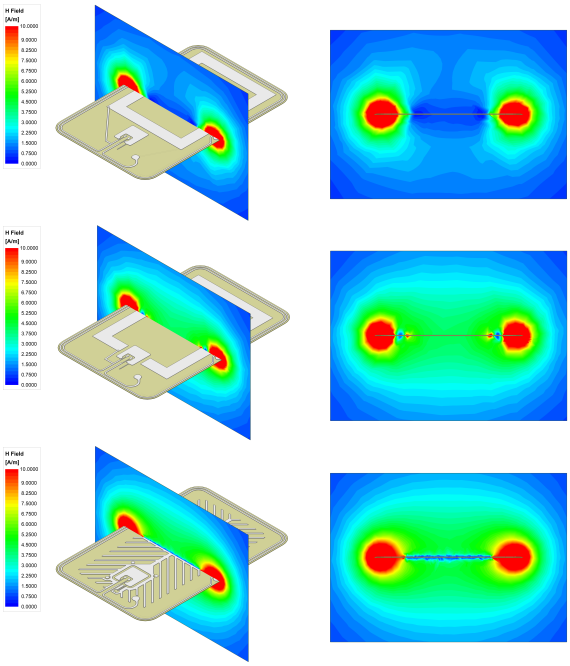


Fig. 11: Magnetic field profile comparison: enclosed frame-shaped design (top), frame-shaped design with slit (middle), and "fishbone" design (bottom)

"Fishbone" design's advantage can also be observed from the resonance curves. Its impedance magnitude was compared to the design with slit (Fig. 12) and the enclosed design

(Fig. 13), respectively. The f_{RES} is at 17.82 MHz, whereas the f_{RES} for the design with slit that has equivalent capacitor area (A_C=98.62 mm^2) lies at 18.88 MHz (Fig. 12). The resonance curve gets shifted towards higher frequencies due to lower effective inductance caused by stronger eddy currents. Also, the Q-factor is reduced due to increased power dissipation. According to (2), in order to tune the f_{RES} of the design with slit back to 17.82 MHz, additional capacitor area would be needed (A_C=108 mm^2). The increase in area would result in further reduction of the Q-factor.

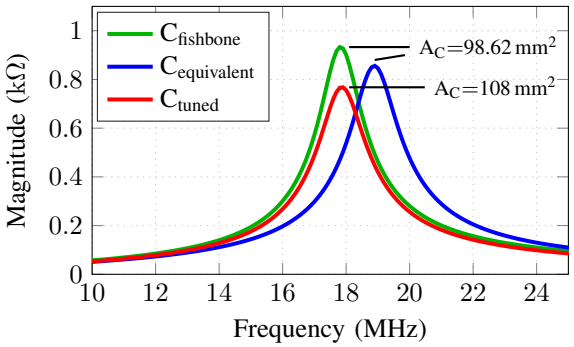


Fig. 12: Resonance curve comparison: "fishbone" design vs. frame-shaped design with slit

As expected, the enclosed frame-shaped design has a more negative effect on the resonance curve (Fig. 13). Here, the curve gets shifted further towards higher frequencies (f_{RES}=19.43 MHz) and has an even lower Q-factor. Consequently, a larger capacitor area (A_C=123 mm^2) would be needed for tuning to 17.82 MHz in this case.

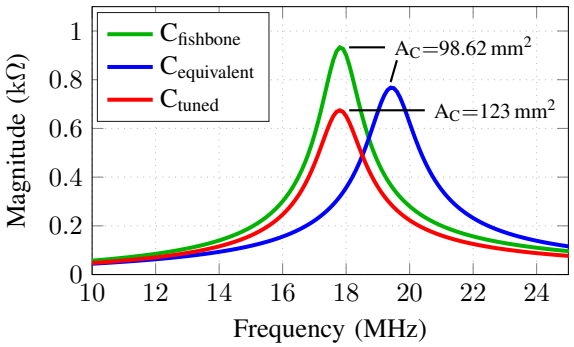


Fig. 13: Resonance curve comparison: "fishbone" design vs. enclosed frame-shaped design

V. Searching for the Optimal Design

Being aware of the potential pitfalls (e.g., compliance to standards, energy transfer, layout geometry effects and material properties), we can apply the simulation framework during the design loop for small payment accessories. Here, we present two different designs and evaluate their performance based on H_{min}. Both are based on the aforementioned SPA module (Fig. 9) and its "fishbone" capacitor. The first design (used to demonstrate the framework in Fig. 4) applies the tuning approach for energy transfer. Its H_{min} was estimated and compared to an actual sample measured on the ISO test PCD assembly (Table Ia). The second design is a geometrically scaled version of the first, about half in size (17.5×13 mm instead of 27.2×17.5 mm), where the "serial first" power matching approach was used to demonstrate the advantage of power matching over tuning (Table Ib). Despite reducing the size by approximately half, Table I shows that almost the same H_{min} performance can be achieved when applying power matching instead of tuning. In addition, Table I shows our simulation framework's excellent accuracy in determining H_{min} when compared to measurements.

TABLE I: Small payment accessories design examples: H_{min}

(a) Tuning

Size (mm)	H_{min} Sim. (A/m)	H_{min} Meas. (A/m)
27.2×17.5	1.88	1.85

(b) Power matching ("serial first")

Size (mm)	H_{min} Sim. (A/m)	H_{min} Meas. (A/m)
17.5×13	1.76	1.80

Beside H_{min}, resonance curves can also give an indication about modeling accuracy. Taking the power matching design as an example, its resonance curve almost completely matches the measurement (Fig. 14). The same values for the equivalent circuit model components can be derived based on both curves (e.g., using curve fitting algorithms), which is a testament to the model's accuracy. The reason why here the impedance magnitude also has a minimum is due to the serial matching capacitor.

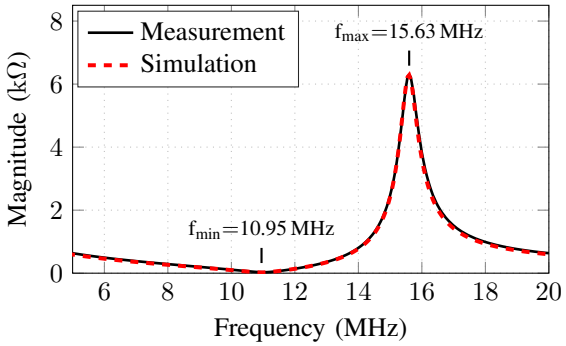


Fig. 14: Resonance curve of the power matching design example: measurement and simulation

Of course, we do not claim that the designs from Table I are optimal, because so many variables comprise the parameter space in which the optimum lies and it is impossible to test out all combinations. However, the framework's advantage is that "parametric sweeps" can be performed much faster than in reality. For example, a design's copper or substrate thickness can be varied in the simulation, followed by evaluating H_{min} of each variation. This process is also more automatized than

evaluating actual fabricated designs and hence saves resources. In the future, we hope to include other performance parameters in the framework, especially the others relevant for compliance to standardization, such as *load modulation amplitude* (LMA) and *PICC loading effect*. We believe there is yet more to be uncovered in the search for optimal design of passive small payment accessories, and that simulation framework is the tool to make this search less challenging.

VI. Conclusion

In this work, we introduced a simulation framework usable for modeling, design and optimization of passive small payment accessories. The ISO test PCD assembly and the PICC antenna design are represented by linear S-parameter models, while the IC is modeled with a custom defined data-based component. Having the ISO test PCD assembly within the framework allows magnetic field strength to be observed, and therefore makes predictions of H_{min} possible, with an accuracy of below 0.1 A/m at H values typical for small payment accessories (around 2 A/m). We then discussed tuning and power matching as the possible approaches for improving the energy transfer between the antenna and the IC of such a device. The choice of layout and materials and its impact on the energy transfer was discussed for an exemplary design (SPA module). Finally, two different designs were modeled and their predicted H_{min} values matched the fabricated samples. Comparison of the two designs showed that performance improvement and miniaturization are possible when the discussed techniques are applied. In conclusion, using the simulation framework to vary the design parameters and assess the H_{min} performance without need for fabrication is proven to aid in the search for the optimal design by reducing the number of redesign loops, i.e., saving costs.

References

[1] "Payment accessories: Design, issuing and best practices," White Paper, Infineon Technologies AG, Nov. 2018.

[2] K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification*, 2nd ed. Wiley, 2003.

[3] EMV payment tokenization, "A guide to use cases," Tech. Rep. Version 1.0, Jun. 2019.

[4] X. Larduinat. (2019, Sep.) EMV tokenization: Digital wallet technology enters the ecommerce space. [Online]. Available: https://blog.gemalto.com/financial-services/2019/09/12/emv-tokenization-digital-wallet-technology-enters-the-ecommerce-space/

[5] J. Grosinger, W. Pachler, and W. Bösch, "Tag size matters: Miniaturized RFID tags to connect smart objects to the internet," *IEEE Microwave Magazine*, vol. 19, no. 6, pp. 101–111, Sep. 2018.

[6] M. Gebhart, M. Wobak, E. Merlin, and C. Chlestil, "Active load modulation for contactless near-field communication," in *2012 IEEE International Conference on RFID-Technologies and Applications (RFID-TA)*, 2012, pp. 228–233.

[7] W. Pachler, J. Grosinger, W. Bösch, G. Holweg, and C. Steffan, "A miniaturized dual band RFID tag," in *2014 IEEE RFID Technology and Applications Conference (RFID-TA)*, Sep. 2014, pp. 228–232.

[8] W. Pachler, J. Grosinger, W. Bösch, P. Greiner, G. Hofer, and G. Holweg, "An on-chip capacitive coupled RFID tag," in *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, April 2014, pp. 3461–3465.

[9] EMV level 1 specifications for payment systems, "EMV contactless interface specification," Tech. Rep. Version 3.0, 2018.

[10] ISO/IEC 7810, "Identification cards - Physical characteristics," Committee identification: ISO/IEC JTC1/SC17/WG1, Tech. Rep., 2003.

[11] ISO/IEC 14443, "Identification cards - Contactless integrated circuit cards - Proximity cards," Committee identification: ISO/IEC JTC1/SC17/WG8, Tech. Rep. Part 1: Physical characteristics, 2016.

[12] ISO/IEC 10373, "Identification cards - Test methods," Committee identification: ISO/IEC JTC1/SC17/WG8, Tech. Rep. Part 6: Proximity cards, 2016.

[13] M. Gebhart, J. Bruckbauer, and M. Gossar, "Chip impedance characterization for contactless proximity personal cards," in *2010 7th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP 2010)*, 2010, pp. 826–830.

[14] W. Lin, "Modellierung, Optimierung und Vermessen von HF-RFID-Transpondern unter Berücksichtigung nichtlinearer Effekte," Ph.D. dissertation, Institut für Hochfrequenztechnik und Funksysteme, Gottfried Wilhelm Leibniz Universität Hannover, 2012.

[15] ISO/IEC 14443, "Identification cards - Contactless integrated circuit cards - Proximity cards," Committee identification: ISO/IEC JTC1/SC17/WG8, Tech. Rep. Part 2: Radio frequency power and signal interface, 2016.

[16] W. Lin, B. Geck, H. Eul, C. Lanschuetzer, and P. Raggam, "A novel method for determining the resonance frequency of PICCs," in *2008 6th International Symposium on Communication Systems, Networks and Digital Signal Processing*, 2008, pp. 311–315.

[17] W. Lin, B. Geck, and H. Eul, "The resonance frequency measurement method of PICCs and the environmental influence," 2007.

[18] M. Pesic, J. Gruber, S. Rampetzreiter, H. Witschnig, and H. Arthaber, "A precise resonance frequency measurement method based on ISO-standardized setups for contactless chip cards," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 29, no. 7, p. e21702, 2019.

[19] "Smart payment accessories (SPA): Fast integration of contactless payment functionality into wearables," Product Brief, Infineon Technologies AG, Oct. 2018.

[20] S. Rampetzreiter, F. Püschner, W. Pachler, H. Witschnig, J. Pohl, and S. M. Wagner, "Antennenmodul, Antennenvorrichtung und Verfahren zum Herstellen eines Antennenmoduls," DE Patent DE102 018 105 383A1, Sep. 12, 2019.

[21] F. Di Paolo, *Networks and Devices Using Planar Transmissions Lines*. Taylor & Francis, 2000.

[22] D. Griffiths, *Introduction to Electrodynamics*. Pearson Education, 2014.

Trust-Provisioning Infrastructure for a Global and Secured UAV Authentication System

Dominic Pirker*†, Thomas Fischer*†, Harald Witschnig†, Christian Steger*
Email: {dominic.pirker, thomas.fischer3, harald.witschnig}@infineon.com, steger@tugraz.at
*Institute for Technical Informatics, Graz University of Technology, Graz, Austria
†Development Center Graz, Infineon Technologies AG, Graz, Austria

***Abstract*—UAVs are gaining momentum in various areas, whether it is in the commercial or private sector. Novel scenarios are extending the seemingly endless list of use cases for this emerging technology. To avoid ungoverned proliferation and abandoned aerial objects, not only regulations but also technical solutions are indispensable. Authentication of a UAV is required to link to the operator and respective competences. Besides appropriate competences, regulations are depending on regional authorities, which demands a studious concept to avoid insular solutions.**

This paper proposes a thought-through infrastructure for a secured and global operative authentication system. First, upcoming regulations are considered for the concept to make the system regulatory compliant. Then, to avoid a patchwork of proclaimed solutions, the system design is based on the principle of delegated authority, which allows the respective authorities to keep control over their domains. Further, to associate UAVs with their operators, a cryptographic link is created during a provisioning process. This link is represented by a certificate, comparable with a conventional driver's license. The system design allows divestment of respective flight permissions, enabled by certificate revocation. Lastly, we constructed a proof-of-concept for the proposed infrastructure solution and compared it to a decentralized approach.

***Index Terms*—UAV, authentication, TLS, certificates, PKI, mDL, HSM, DNS**

I. Introduction

Registration and subsequent identification of Unmanned Aerial Vehicles (UAVs) is getting essential, since the UAV market is highly dynamic and will heavily increase in the upcoming years. For instance, in Germany the market will grow six-fold in the next decade, from 500 Million Euro to 3 Billion Euro [1]. The number of operational UAVs will increase to almost 1 Million by 2023 [1].

Considering the tremendous growth - infrastructure, services, and procedures have to provide safe Unmanned Aerial System (UAS) operations and support their integration into the aviation systems [2]. A general problem are the region-depended regulations. Therefore, a concept for a global operational system is required to enable automatic position detection as well as intermediate application of region-depended regulations. Fig. 1 depicts a global and secured UAV authentication system, consisting of a flight control and a UAV. On top of the protected communication channel, which is supported by an Hardware Security Module (HSM) on the UAV, the flight information is transmitted. A problem with UAV identification systems is the missing link between the UAV and its operator. Therefore, the system depicted in Fig. 1 needs to be extended.

The main contributions of this work are:

- Proposal of an infrastructure for a global aviation system, based on the global and secured UAV authentication system depicted in Fig. 1 and proposed in [3].
- Designing the certificate deployment process to be globally operational.
- Digital licensing based on a cryptographic link between the UAV and the corresponding operator.
- Evaluation of the proposed infrastructure system and analysis of potential weaknesses.

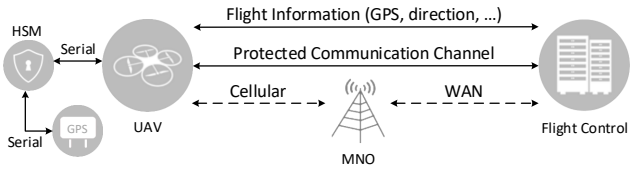


Fig. 1. Connection overview between flight control and UAV including an HSM (adopted from [3])

II. State of the Art

A. Regulatory Framework

The European Union Aviation Safety Agency (EASA) released common rules on UAS to secure UAS operations [4]. UAS describes the entire system consisting of the UAV, the ground control station or pilot, and other components such as camera or Global Positioning System (GPS) receiver. Two related documents have been published by the EASA: *Commission Delegated Regulation (EU) 2019/945* [5] and *Commission Implementing Regulation (EU) 2019/947* [2]. The *Delegated Regulation* describes requirements for designing and manufacturing UAS to allow operation based on rules and conditions defined in *Implementing Regulation* [5]. The latter lays down provisions for the operation of UAVs, including personnel and organizational usage [2]. In addition, industry should not diminish in agility, innovation, and continuous growth, while implementing the regulations.

These regulations include requirements for implementation of three services of the *U-space* system, that is a set of services and procedures to support safe, efficient, and secured access to the airspace for a large number of UAVs [6]. The operational

requirements are geo-awareness, remote identification, and registration of the operator and UAV. Even tough the regulations are defined by the European Union (EU), EASA member states will still remain flexible in the context of defining zones or additional requirements.

B. Digital Driver's License

Digital driver's license is often referred to as mobile driver's license. This kind of driver's license is different from electronic driver's licenses. The latter is basically the traditional driver's license in ID card format, equipped with a security controller to protect personal data such as biometric data [7]. Electronic driver's licenses are, as electronic passports, already available in many countries, whereas digital driver's licenses are for now only in testing and pilot phases [8]. National Institute of Standards and Technology (NIST) is pushing a pilot project in the US together with Gemalto to verify the technical feasibility of digital driver's licenses [9]. Further, the International Organization for Standardization (ISO) has a dedicated working group to tackle the emerging trend towards digital driver's licenses (ISO/IEC JTC 1/SC 17/WG 10). In this regard, the high impact of drones is pointed out by the foundation of a dedicated working group, having drones, their licensing, and further their operator's identity as major subjects (ISO/IEC JTC 1/SC 17/WG 12).

C. Public Key Infrastructure

Public Key Infrastructure (PKI) is a well-established, widely used and centralized mechanism to enable trust, with the key elements: confidentiality, authenticity, integrity, and non-repudiation [10]. In [11], a PKI infrastructure for a large-scale Internet-based healthcare network is proposed to provide security for connecting a wide-spread spectrum of geographically distributed units. The authors from [11], adopted the traditional hierarchical PKI trust model to enable compartmentalization of different responsibilities. This is also considered in the design of the global aviation infrastructure system, proposed in this work.

Main responsibilities of PKI systems are certificate issuing, certificate deployment, and certificate validation (typically X.509 certificates). Based on public-key cryptography, messages sent via an insecure network can be digitally signed and encrypted. To provide the affiliation of public keys, digital certificates are used. The primary party of a hierarchical PKI system is the Certificate Authority (CA), also acting as a registration and validation authority simultaneously. Web of trust is a different approach for public authentication, which is based on *OpenPGP* and standardized in RFC 4880 [12].

D. DNS Namespace

The Domain Name System (DNS) is a hierarchical naming system that links IP addresses to domain names. These domains exist in various levels and are connected in a hierarchical tree structure [13]. Example: *maps.google.com*; *"com"* is the top-level domain; *"google"* is a sub-domain and *"maps"* is a lower-level sub-domain. With few exceptions, the domains are associated to regions (e.g. *"at"*, *"de"*, or *"us"*). This regional and hierarchical approach is chosen for the concept of the global lookup service proposed in this paper.

III. The Global Aviation Infrastructure System

The system we proposed in [3], describes a *Global and Secured UAV Authentication System based on Hardware-Security*, that uses the Transport Layer Security (TLS) protocol, supported by an HSM (depicted in Fig. 1). This system requires an established PKI infrastructure. Based on that, the certificate provisioning procedure is performed. This procedure is split into UAV (client) and server authentication. The UAV authentication part is associated with the UAV itself, the UAV manufacturer CA, and the smart remote control (e.g. smartphone), depicted in Fig. 3. The server authentication part is associated with flight control servers, regional CAs, and a global aviation authority lookup service. The server authentication has similarities to the DNS lookup service and is depicted in Fig. 2.

A. Requirements

The main requirements for the infrastructure of the global aviation system including the regulative requirements from the the *Delegated* [5] and *Implemented Regulation* [2] are:

- *Global availability* is a necessity to avoid insular solutions. This includes capability to comply with regulations in respective regions, as well as allowing authorities to keep control over their flight zones.
- *Authentication*, not only identification, which implies the necessity to proof the identity (not just object classification as radar-based systems [14]).
- Regulative requirements:
  - *Geo-awareness* to allow the implementation of no-fly zones.
  - *Remote Identification* to know the operator during flying.
  - *Registration* of *Operator* (Pilot) to allow later identification and verification of possibly necessary proof of knowledge and competences.
  - *Registration* of *UAV* to allow classification and verification of certified hardware.

An additional requirement that was defined during the research process, is the creation of the cryptographic link between the UAV and the operator while authenticating against the flight control server, explained in more detail in Section III-C2.

B. Goal

A UAV that is switched on, must authenticate itself and its operator against a flight control. As regulations may be regional dependent, the UAV must be able to choose the location corresponding flight control server. Therefore, a global lookup service is required. To achieve this high level of scalability, the certificate provisioning process has to be separated into two parts, the preliminary steps and the operational steps.

C. Preliminary Steps

These steps are separated into server and client setup. Each paragraph explains the respective steps for provisioning the certificates before the actual UAV and operator authentication against the flight control server happens.

1) Server Authentication: For this infrastructure, two different types of servers exist, the global lookup server and at least one flight control server. Reasons for multiple flight control servers may include the amount of UAVs, size of the region, and redundancy. Both, the global lookup server and the flight control server must support server authentication. Fig. 2 depicts the steps for the server authentication. In the following figures the preliminary and operational steps are visualized with dashed and solid lines, respectively.

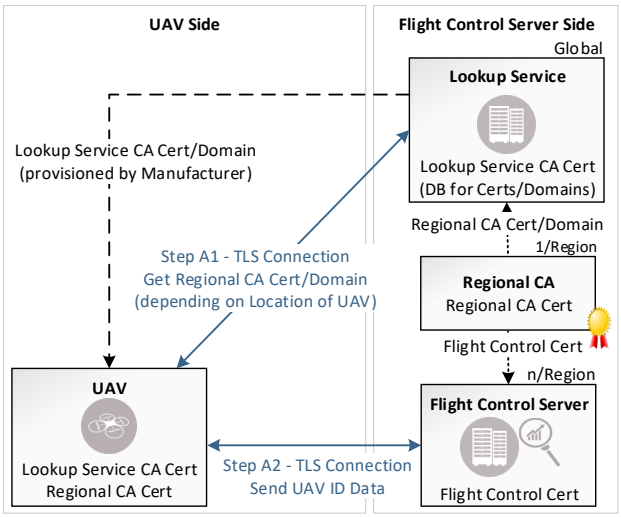


Fig. 2. Infrastructure and certificate provisioning process for server authentication

For authentication of the global lookup service against a UAV, the lookup service certificate must be stored on the UAV, for instance in the protected storage of an HSM. For security reasons, the lookup service certificate should be stored in read-only memory (ROM) to avoid tampering with the certificate. Additionally to the lookup service certificate, the IP address or domain of the global lookup service must be stored in the UAV's memory during the manufacturing process, to be able to request the regional authority certificate and domain later. For the current approach, client authentication against the global lookup service is not required, because it is a public service.

For authentication of the flight control server against a UAV, the regional CA issues and provisions a certificate to each flight control server. The regional CA certificate is stored, together with the respective IP address or domain, at the global lookup service.

2) Client Authentication: Client authentication is required for two reasons, authenticating the UAV against the remote control, and most essentially authenticating the UAV and its operator against the flight control server. To enable this, several steps are necessary as depicted in Fig. 3.

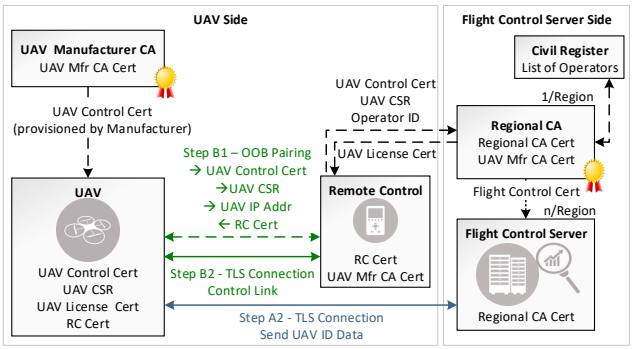


Fig. 3. Infrastructure and certificate provisioning process for client authentication

First, the UAV needs to be paired with the remote control at first use, labeled as *Step B1* in Fig. 3. During pairing, the certificates are exchanged between UAV and remote control, and the IP address of the UAV is stored on the remote control. For this, an Out-of-Band (OOB) pairing method has to be implemented. Specifying the exact method is out of scope, but Near Field Communication (NFC) or a method described in [15] can be used. The RC certificate is a self-signed certificate, and is generated on the remote control.

After the pairing is complete, the secured channel can be used to transmit details for generating the UAV license certificate to securely connect and authenticate to the flight control server. The details include the UAV control certificate and the UAV Certificate Signing Request (CSR).

Preliminary, the UAV manufacturer CA issues a UAV control certificate for each UAV and stores it in the HSM's protected storage. This certificate is used by the UAV to authenticate itself against the remote control, that holds the UAV manufacturer certificate for validation. The deployment of the UAV manufacturer certificate is out of scope, but one solution is to deliver it along with the mobile application for remote control.

The next steps are designed to cryptographically link the UAV and the operator for authentication against the flight control server. This step is necessary, since TLS is designed to utilize exactly one certificate per peer for connection establishment. Alternatively, a second certificate can be sent at the application layer, but this would mix the application with the security layer.

The regional CA is in charge of this linking procedure. Therefore, the UAV control certificate and the UAV CSR, together with a personal identifier of the operator, are required. The CSR is generated by the UAV itself. Specific UAV related information (e.g. model, weight, etc.) are extracted from the UAV control certificate, which is provisioned by the UAV manufacturer. Then, the CSR is generated with this as input and signed with the private key, that is stored in the UAV's HSM. The channel establishment between the remote control and the regional authority is out of scope of this work. Possible solutions include a web API or a separate smartphone application.

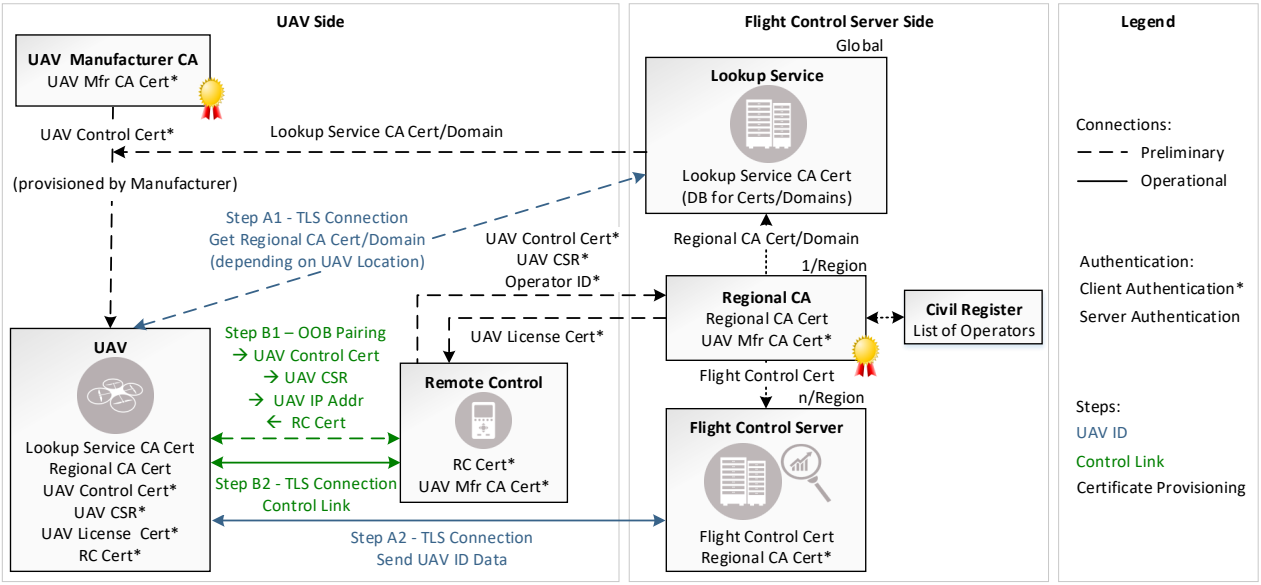


Fig. 4. Infrastructure and provisioning process for the global and secured UAV authentication system

At the regional authority, the UAV control certificate is used to check, that the UAV to be registered, is from an accredited UAV manufacturer. The signature of the CSR is validated with the public key of the UAV's control certificate to check, the requester possesses the UAV, for which a certificate is requested. Next, the ID of the operator and the UAV details are processed and validated for required competences, based on a civil register. If so, the UAV CSR is extended with operator ID data and signed by the regional CA, which results in the UAV license certificate.

The UAV license certificate is sent to the remote control first, then it is sent via the TLS connection to the UAV, where it is finally used for authentication to the flight control server.

D. Operational Steps

After the preliminary steps are complete, the system is ready for operation. First, this includes establishing a protected channel between UAV and remote control labeled as *Step B2*. Second, establishment of a protected channel between UAV and flight control server (*Step A2*). The latter, requires a protected communication channel to the global lookup service beforehand, in order to request the location corresponding flight control server's domain.

The TLS protected control link is established with the UAV control certificate validated with the UAV manufacturer certificate, and the RC certificate validated by the remote control.

Following the boot process of a UAV, a TLS connection to the global lookup service, labeled as *Step A1* in Fig. 2, is established. The global lookup service certificate, stored in the UAV's memory, is used for server authentication.

Using the protected connection, domain, and certificate of the location-dependent regional authority are requested. The required UAV location information can either be provided as GPS coordinates by the UAV, or the lookup service can locate the UAV based on its IP address if the Mobile Network Operator (MNO) provides location specific addresses. Another possibility to get the location information of the UAV, is to extract this information from the connected LTE cell as described in [16].

The TLS connection to the lookup service is closed and based on the recent obtained domain, a TLS protected communication channel from the UAV to the flight control server, labeled as *Step A2* in Fig. 2, is established. The identity of the flight control server is validated by the regional CA certificate retrieved in *Step A1*. The authentication of the UAV and its operator is done with the UAV license certificate. This certificate is validated with the UAV manufacturer certificate stored on the flight control server.

In Fig. 4 both, the preliminary and operational steps, together with all involved parties are depicted.

E. CA Hierarchy

The proposed infrastructure requires three different root CAs, the UAV manufacturer CA, the regional CA, and the global lookup service CA. As depicted in Fig. 5, these root CAs are independent. In the described hierarchy, the root CA is at the same time the issuing CA.

The UAV control certificate is issued by the manufacturer CA. The regional CA is issuing the UAV license certificates and the flight control certificates. The lookup service CA certificate is used for server authentication, and does not issue additional certificates.

F. Permission Revocation

Divestment of respective flight permissions of specific operators is allowed due to the design of the system. In PKIs, a certificate is expected to be valid for the entire period,

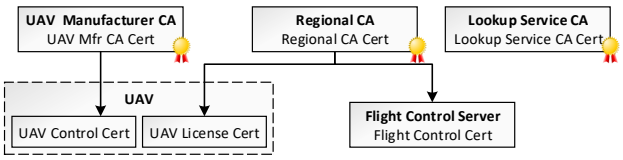


Fig. 5. Certificate authority hierarchy for the global UAV authentication system

set during issuing. Typically, scenarios such as changes in association between subject (e.g. UAV) and CA, or suspected compromise of the corresponding private key, can cause prior invalidation of the certificate [17]. The invalidation is enabled by adding identifier of the corresponding certificate to the Certificate Revocation List (CRL), which is issued periodically to distribution points which are linked within the certificate. One problem with CRL is the delay, caused by the periodic update. Another attack against this revocation mechanism is a Denial-of-Service attack against the distribution points, which would freeze the current status of the CRL, and therefore new permission revocation are affected. A CRL allows two different states of revocation: Revoked (irreversible) and hold (reversible).

The proposed system enables revocation of flight permissions by adding the respective UAV license certificate to the CRL. This is done by the flight control server, based on regional defined regulations. An example could be restricted behavior of the operator, such as flying over highly critical areas (e.g. airports, power plants).

IV. Proof-of-Concept

For the proof-of-concept, a manual step-by-step execution of the certificate provisioning process is performed. For the hardware, we build upon the setup used in [3]. The setup consists of a modular UAV equipped with a Raspberry Pi Zero which was extended with an I2C-connected HSM, and a Raspberry Pi 3, running as a flight control server. Both are operating with Raspbian Stretch. For certificate issuing, the *OpenSSL* toolkit is used, which is publicly available and licensed for commercial and non-commercial usage [18].

A self-signed CA certificate for the UAV manufacturer is generated. With that, a UAV control certificate is issued and stored at the UAV's HSM. Next, a CSR is generated with support of the HSM, which stores the UAV's private key in protected memory. As input for the subject field of the CSR, dummy data is used to simulate UAV related details, such as model or weight. In the proof-of-concept, remote control, flight control, and regional authority are hosted on the same physical system, but distinct TLS channels are established. The certificates used for establishing the control link, are the RC certificate, validated by the remote control, and the UAV control certificate, validated with the UAV manufacturer certificate. The OOB pairing is performed manually by putting the certificates on the corresponding devices.

On the flight control server, a regional CA certificate is generated and a flight control server certificate is issued. To issue the UAV license certificate, which is linking the operator with the UAV, first the UAV CSR is validated with the UAV manufacturer CA. Then, the CSR is extended with dummy attributes (representing operator details) and the certificate is issued with the regional CA certificate and the corresponding key. Then, this certificate is sent to the UAV, where it is used for authentication and protected connection establishment against the flight control server. The certificate's validity is checked with the regional CA certificate.

V. Evaluation

A. Digital Licensing

One key concept of the proposed infrastructure for a global and secured UAV authentication system is the UAV licensing. It is comparable with a digital driver's license, since both licenses are stored on an embedded device and not on chip cards representing a document. The difference is, within the pilot projects, digital driver's licenses are stored on the mobile phone [9], whereas in this concept, the license is stored at the vehicle (UAV in this case). The license is represented by an X.509 certificate (UAV license certificate) and is linked to a specific UAV. Comparing to traditional licensing use cases, for instance car driver's license, the operator is not always in the same location as the UAV, and therefore a link between the vehicle and the operator is mandatory.

B. DNS Similarities

The proposed global lookup service has strong similarities to DNS. Both are a hierarchical mapping of a dynamic database scattered globally [19]. As DNSSEC, the trust relationship has to be built from the root, which is corresponding to the global lookup service within our proposed concept. Trust is established by verifying the lookup server's identity during the TLS handshake, with support of the lookup service CA certificate, stored in the UAV's HSM during manufacturing. The location of the UAV is comparable with the country code within DNS. This concept design was chosen, because the concept as DNS is utilizing, is well established and widely used for providing global available services [19].

C. Concept Evaluation

The global lookup service proposed in this work, brings a major advantage. Due to the fact that in *Step A1* of the provisioning process, the location-dependent domain, respectively the according flight control server certificate, is fetched from the global lookup service, a UAV always connects to the correct, location-corresponding server. Using this measure, regional regulations can be defined by the respective authorities, even tough a global system is used.

One drawback of the proposed infrastructure concept is the potential single point of failure which applies for the global lookup service and the flight control servers. If one of those fails or is attacked, the system might become unavailable. If an attacker manages to retrieve the private key corresponding

to the certificate of a server, trust in the entire system is compromised [20].

A countermeasure is to implement redundancy, as briefly described for the flight control servers in Section III-C1. A comparable approach can be implemented for the global lookup service. Therefore, IP or domain to secondary or even tertiary global lookup service can be provisioned during UAV manufacturing. Again, DNS implements a similar approach, where a client can store the IP addresses of multiple DNS servers. Alternatively, a decentralized concept, for instance based on blockchain, is a conceivable solution. Blockchain, a decentralized ledger of transaction, solves the problem of single point of failure, compared to digital certificate systems [20].

VI. Conclusion and Future Work

In this work we proposed an infrastructure and trust provisioning process for a global operative and secured UAV authentication system, that allows authentication of both, the UAV and the corresponding operator. The design allows permission divestment, in cases such violation of no-fly zones is detected. Additionally, region-dependent regulations are respected, which is supported by the implementation of the global lookup service and the corresponding flight control servers.

Future work will further investigate on the weaknesses of the proposed system. Alternative solution for the given problem statement, for instance based on blockchain, as mentioned in the evaluation, will be researched. A promising approach is to design a hybrid solution that combines the advantages of hierarchical and decentralized solutions.

VII. Acknowledgment

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826610. The JU receives support from the European Unions Horizon 2020 research and innovation programme and Spain, Austria, Belgium, Czech Republic, France, Italy, Latvia, Netherlands.

References

[1] Verband Unbemannte Luftfahrt (VUL), "Analysis of the German Dronemarket," https://www.bdl.aero/wp-content/uploads/2019/02/VUL-Markststudie_Deutsch_final.pdf, [Online; accessed 2020-03-14].

[2] Council of European Union, "Commission implementing regulation (eu) 2019/947," https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R0947&from=EN, 2019, [Online; accessed 2019-12-04].

[3] Dominic Pirker, *Design and Implementation of a Global and Secured Drone Identification System with Hardware-Based Security*. TU Graz, 2019.

[4] EASA, "EU wide rules on drones published," https://www.easa.europa.eu/newsroom-and-events/news/eu-wide-rules-drones-published, [Online; accessed 2019-10-21].

[5] Council of European Union, "Commission delegated regulation (eu) 2019/945," https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R0945&from=EN, 2019, [Online; accessed 2019-12-04].

[6] SESAR Joint Undertaking, "U-space blueprint," https://www.sesarju.eu/sites/default/files/documents/reports/U-space%20Blueprint%20brochure%20final.PDF, 2017, [Online; accessed 2020-01-28].

[7] A. A. of Motor Vehicle Administrators (AAMVA), "Mobile Driver's License," https://www.aamva.org/FunctionalNeedsWhitepaper-9/, Whitepaper, 2016, [Online; accessed 2020-03-24].

[8] R. T. Raj, S. Sanjay, and S. Sivakumar, "Digital License mv," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016, pp. 1277–1280.

[9] NIST, "Digital identity for individuals," https://www.nist.gov/itl/tig/digital-identity-individuals, [Online; accessed 2020-04-20].

[10] B. Rajendran, "Evolution of PKI ecosystem," in *2017 International Conference on Public Key Infrastructure and its Applications (PKIA)*, Nov 2017, pp. 9–10.

[11] G. Mantas, D. Lymberopoulos, and N. Komninos, "PKI Security in Large-Scale Healthcare Networks," *Journal of medical systems*, vol. 36, pp. 1107–16, 09 2010.

[12] e. a. J. Callas, "OpenPGP Message Format," Internet Requests for Comments, RFC Editor, RFC 4880, November 2007. [Online]. Available: https://tools.ietf.org/html/rfc4880

[13] P. Satam, H. Alipour, Y. Al-Nashif, and S. Hariri, "DNS-IDS: Securing DNS in the Cloud Era," in *2015 International Conference on Cloud and Autonomic Computing*, Sep. 2015, pp. 296–301.

[14] M. Jian, Z. Lu, and V. C. Chen, "Drone detection and tracking based on phase-interferometric Doppler radar," in *2018 IEEE Radar Conference (RadarConf18)*, April 2018, pp. 1146–1149.

[15] H. Nakajima, S. Suzuki, T. Tokunaga, K. Tanaka, Y. Miyazaki, K. Maruyama, and O. Nakamura, "Out-of-band authentication protocol for digital signage and smartphone interaction," in *2016 IEEE 5th Global Conference on Consumer Electronics*, Oct 2016, pp. 1–2.

[16] Sven Fischer, *Observed Time Difference Of Arrival (OTDOA) Positioning in 3GPP LTE*, 1st ed. Qualcomm, 2014.

[17] e. a. Housley, "Internet X.509 Public Key Infrastructure Certificate and CRL Profile," Internet Requests for Comments, RFC Editor, RFC 2459, January 1999. [Online]. Available: https://tools.ietf.org/html/rfc2459

[18] OpenSSL Software Foundation, "OpenSSL," https://www.openssl.org/, [Online; accessed 2020-04-20].

[19] M. H. Jalalzai, W. B. Shahid, and M. M. W. Iqbal, "DNS security challenges and best practices to deploy secure DNS with digital signatures," in *2015 12th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2015, pp. 280–285.

[20] R. Wang, J. He, C. Liu, Q. Li, W. Tsai, and E. Deng, "A Privacy-Aware PKI System Based on Permissioned Blockchains," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Nov 2018, pp. 928–931.

Low Noise IQ Generation Employed in an Active Vector Modulator for 5G Ka-Band Beam Forming Transceivers

B. Mesgari, and H. Zimmermann

Institute of Electrodynamics, Microwave and Circuit Engineering,
Vienna University of Technology, Gußhausstraße 25/E354, A-1040 Wien, Austria
E-mail:{baset.mesgari | horst.zimmermann}@tuwien.ac.at

***Abstract*— This paper investigates a compact, low noise differential in-phase (I) and quadrature signal (Q) generator. The proposed circuit utilizes a pair of mutual inductors to generate 90° phase shift which results in the same voltage gain for both I and Q paths. while it is in contrast with an RC-CR structure where the same gain for IQ paths cannot be achieved with a proper 90° phase shift simultaneously. The simulation results in a 130 nm SiGe-BiCMOS technology, indicate that a 3-dB bandwidth of 1.54 GHz around 28 GHz center frequency is achieved. Employing the proposed structure in a low-IF receiver as part of the phase-shifting block has indicated a voltage gain of 30 dB for the desired signal while 40 dB rejection is provided at 22.275 GHz for the image signal. NF and S_{11} are less than 3.5 dB and -17 dB, respectively, for the entire band of interest while the receiver consumes 8 mA from a 2V power supply. The phase error is less than 5° and the gain variation error is 0.3 ~ 0.5 dB. The input IP3 for the rest of the receiver chain is about -27 dBm.**

Keywords—5G beamforming, active vector phase shifter (PS), IQ signal generator, phase and gain error, mm-wave

I. INTRODUCTION

Fifth-generation (5G) wireless communication due to the unused spectrum at millimeter-wave (mm-wave) is an emerging network that can increase data rate significantly. Especially, the 28-GHz frequency band can be an interesting choice for implementing and realizing high-performance hardware in an integrated circuit fabrication process like CMOS and SiGe-BiCMOS [1]-[3]. In a 5G transceivers, phased array antennas enable directional and spatial filtering, thereby enhancing the link reliability as well as the interference signals suppression [4]. As shown in Fig.1, spatial filtering in a 5G receiver is accomplished through a group of antennas in which the relative phase of the received signals can be aligned using a phase shifter (PS) in such a way that each radiation pattern is strengthened in a wanted direction and relatively attenuated in other directions [4]-[7]. Thus, PSs in this structure play a critical role as far as they should define a rather precise phase while providing noise figure (NF), linearity and power consumption as proper as possible in a frequency band of interest [6], [8]. To realize a PS, different approaches and topologies have been reported in the literature which can be classified into two general strategies namely active phase shifting and passive phase shifting. Based on Fig.1, since the PS is placed after LNA, therefore their components loss can cause NF degradation and signal attenuation remarkably. Moreover, a passive PS is frequently composed by cascade number of the tunable T-

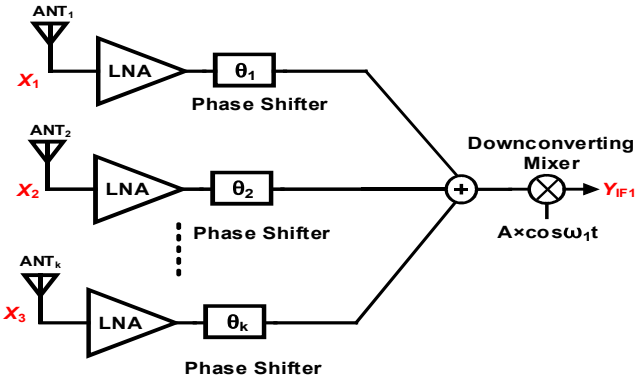


Fig. 1. A low-IF phased array receiver, employing the spatial filtering.

networks for each path which result in taking a large die area due to utilizing inductors. If the occupied area and NF are essential parameters for a beamforming receiver, employing a passive PS would not be the best choice, in contrast, in this case, active PS can be the best one. Fig. 2 depicts the circuit diagram of a vector modulator that belongs to the active PS category. The desired phase shift can be adjusted by the current signal gain of Q path $G_{M\text{-}Q}$, divided by the current signal gain of the I path $G_{M\text{-}I}$. As presented in this figure the amplified differential signal at the output of the LNA is turned to a perpendicular signal (V_I, V_Q) by using a quadrature signal generator. The phase resolution of the vector modulator depends on the accuracy of the quadrature signal generator block. As a result, a precise design of this block is vital for a suitable performance of the entire receiver. To obtain a 90° phase shift, a large number of popular techniques have been developed. The most widespread method is using an RC–CR poly-phase filter (PPF) which is illustrated in Fig. 3(a) [9]. Although, PPF is the best choice considering the occupied area efficiency, for a mm-wave operating frequency a large amount of signal attenuation and capacitive loading in a PPF are not negligible and they can considerably degrade the phase resolution of the vector modulator. To overcome and address the mentioned limitations in the design of a PS and provide a relatively precise phase resolution for the vector modulator, in this paper a mutually coupled all-pass network (MCAN) is presented as its circuit diagram has been illustrated in Fig. 3(b). The proposed circuit employs a pair of mutual inductors to generate a 90° phase shift which can lead to an equal signal gain for both I and Q paths, it is in contrast with PPF structure where the gain

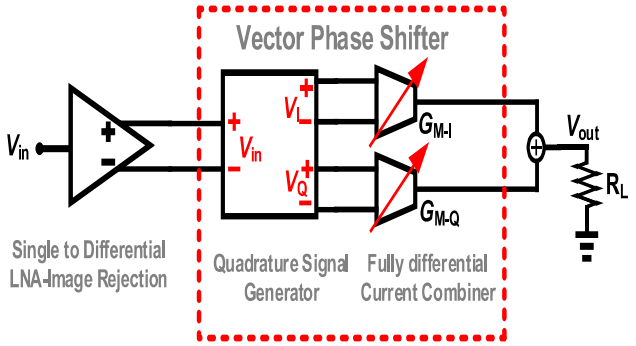


Fig. 2. An active phase shifter based on quadrature signals (IQ) combining.

of the *I* path is always less than the gain of the *Q* path [9]. In addition, our proposed structure has a second-order transfer function which can introduce a degree of freedom to define signal loss and operating frequency independently. MCAN can also be designed more area-efficient in comparison with one which was used in [7] as far as using mutual inductors.

The paper is organized as follows. Section II describes the proposed MCAN topology and explains how much phase and gain can be achieved. Based on the equations and method in this section, phase error in comparison with 90° due to circuit elements will be discussed and finally usage of MCAN in a low-IF receiver as part of a vector modulator will be explained in details. Section III presents simulations result of a low-IF receiver employing MCAN in a 0.13 µm SiGe-BiCMOS technology. In section IV a conclusion is provided.

II. Proposed 90° Phase Synthesis and Design

A. Principal Operation of a Vector Modulator

Fig. 2 exhibits the conceptual block diagram of a vector modulator PS which operates based on a phase interpolation manner by dividing two appropriately scaled quadrature signals. Considering steady-state sinusoidal analysis, V_{out} in Fig. 2 can be calculated using (1).

$$V_{out}(j\omega) \cong R_L\left(G_{M-I} \times V_I(j\omega) + G_{M-Q} \times V_Q(j\omega)\right) \quad (1)$$

Here G_{M-I} and G_{M-Q} are the transconductances of the *I* and *Q* paths. $V_I(j\omega)$ and $V_Q(j\omega)$ is obtained in equation (2), in which *j* is considered by its property $j^2 = -1$. Assuming that the phase shift of the LNA is zero, the resulting phase shift of the output voltage in comparison with the input signal is found in (3). Fig.4 evidently depicts that θ_{out} can be adjusted by properly choosing the transconductance of G_{M-I} and G_{M-Q}. For example, if θ_{out}= 45°, G_{M-Q} is equal to G_{M-I}.

$$V_I(j\omega) = \pm jV_Q(j\omega) = A_{LNA}(j\omega) \times V_{in}(j\omega) \quad (2)$$

Here $A_{LNA}(j\omega)$ is the gain of the single to differential LNA.

$$\theta_{out} = \pm a\,\mathrm{rctan}\left(\frac{G_{M-Q}}{G_{M-I}}\right) \quad (3)$$

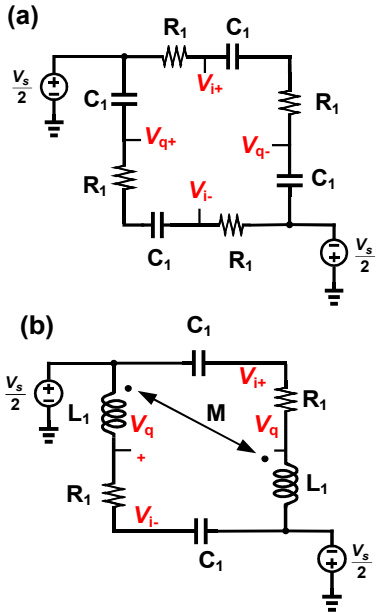


Fig. 3. Circuit diagram illustration of the RC-CR (a) and of the proposed MCAN topology (b).

B. MCAN Analysis and Design

In this section, the gain and phase of the proposed all-pass network is analyzed in details. This analysis leads to simplified equations of the phase and gain response of the proposed topology. By applying the Kirchhoff's voltage and current laws for the network shown in Fig. 3(b), and combining their results, the differential voltage transfer-functions of (V_{di}=V_{i+}-V_{i-}) and (V_{qi}=V_{q+}-V_{q-}) in respect to the input voltage are summarized in (3), (4). In these equations, *M* expresses the mutual inductance between the inductors. By defining the angular frequency $\omega_0 = 1/\sqrt{C_1(L_1+M)}$ and the network time constant $\tau_0 = R_1C_1$ and also substituting $\omega = 2\pi f$ in equations (3) and (4) and doing some algebraic simplifications, the absolute value of $H_I(j\omega)$, $H_Q(j\omega)$ as well as the phase deference between *I* and *Q* paths, which is defined as θ_{DT}, are calculated in (5) and (6), respectively.

$$H_I(j\omega) = \frac{V_{di}}{V_s}(j\omega) = \frac{-\left(1+(L_1+M)C_1\omega^2\right) + j\omega R_1C_1}{\left(1-(L_1+M)C_1\omega^2\right) + j\omega R_1C_1} \quad (3)$$

$$H_Q(j\omega) = \frac{V_{dq}}{V_s}(j\omega) = \frac{\left(1+(L_1+M)C_1\omega^2\right) + j\omega R_1C_1}{\left(1-(L_1+M)C_1\omega^2\right) + j\omega R_1C_1} \quad (4)$$

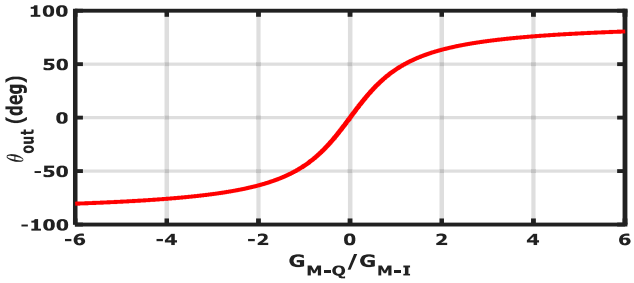


Fig. 4. Phase response based on equation (6).

Assuming that $f=f_0$ and $2\pi f_0\tau_0=2$, results in a 90° phase shift between I and Q paths which can show the efficiency of our proposed all-pass network. The difference between θ_{DT} from 90 degrees is usually called phase error θ_{Er}, which might be due to the mismatch between the elements of the MCAN. A simulation has been done in Fig.5 in order to investigate how much phase error will occure if there is a relatively large deviation from nominal value of C_1, L_1 and R_1. The simulation has been accomplished at the target frequency of 28 GHz, which leads to choosing a set of values such as C_1=100fF, L_1=215pH, M=110pH and R_1=113Ω. This simulation shows that the MCAN is an area efficient topology which can be used in the *Ka* band while the mismatch of their components does not introduce a critical error.

$$|H_I|=|H_Q|=\sqrt{\frac{\left(1+\frac{f^2}{f_0^2}\right)^2+(2\pi f\tau_0)^2}{\left(1-\frac{f^2}{f_0^2}\right)^2+(2\pi f\tau_0)^2}} \tag{5}$$

$$\theta_{DT}=2\times a\,\mathrm{rctan}\left(\frac{2\pi f\tau_0}{\left(1+\frac{f^2}{f_0^2}\right)}\right) \tag{6}$$

C. A Low IF 5G Receiver Using MCAN Phase Shifter

To examine the influence of our proposed IQ generation method, on the performance of the entire RF chain, a low-IF receiver has been chosen. In this test, MCAN has been placed as a quadrature signal generator (see Fig. 2). For implementing the LNA, an image rejection single to differential LNA which is shown in Fig. 6 has been designed at a target frequency of 28 GHz with the bandwidth of 1.54 GHz. L_{im} and C_{im} have also been employed to eliminate the image signal at the frequency of 22.275 GHz, thus the first IF signal can be obtained at 2.8625 GHz (Y_{IF1} in Fig.1). As illustrated in Fig. 3(b), MCAN needs to have a differential voltage at its input, hence the LNA has to provide a differential signal. Consequently, to generate the mentioned differential signal at the output of the LNA, an integrated transformer has been employed with a coupling coefficient of k=0.7.

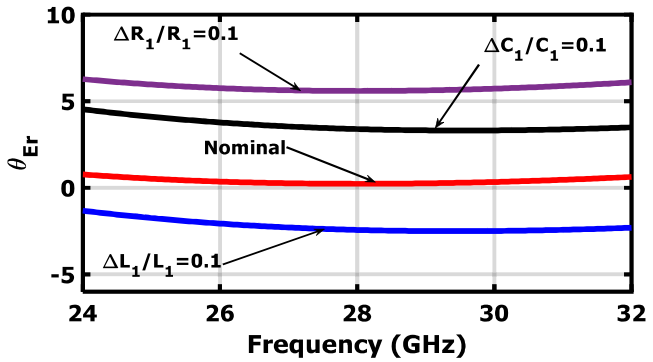


Fig. 5. Phase error for different parameters, which indicate that the MCAN is not a sensitive topology to mismatch of its components.

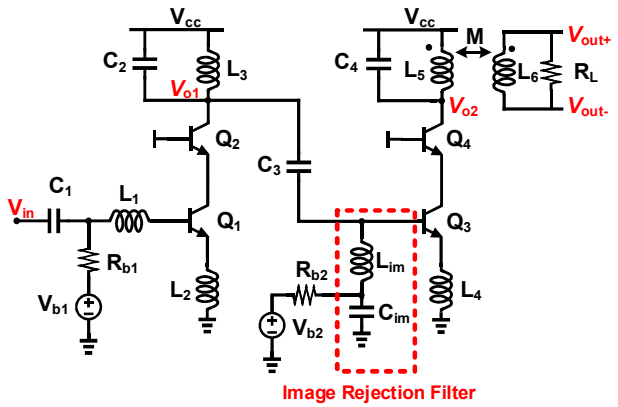


Fig. 6. Image rejection single to differential LNA using an integrated transformer.

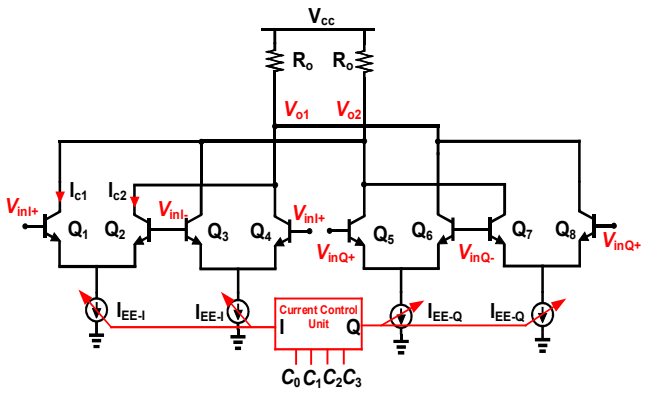


Fig. 7. Transconductance stage (current combining) for I and Q path.

Unity gain differential buffers have been placed at the outputs of the MCAN in order to minimize the loading effect on the performance of generating quadrature signals. A sign sensitive current combiner follows the MCAN, which is presented in Fig. 7. As highlighted in this figure a current DAC is able to provide a digital code ($C_0C_1C_2C_3$) which can be used to control the I and Q current branches. Based on Fig. 4, the proper digital code can define, which transistor pairs should generate how much current with respect to the required phase shift.

D. LNA and the Current Combining Circuit's Analysis

Providing a good matching at the input port (V_{in}) shown in Fig. 6, Z_{in} which expresses input impedance, should stay near to 50Ω in the frequency range of interest. Z_{in} can be calculated from (7). At $\omega_0=1/\sqrt{(L_1+L_2)C_{be1}}$, Z_{in} is $(g_{m1}L_2/C_{be1})$ and should be equal to Rs which represents the antenna resistance. For the sake of simplicity, we assume that C_{bc} and r_o of all transistors are relatively affectless and they can be neglected in our analysis. In addition, as far as C_1 and C_3 are the ac coupling capacitors, they show quite small impedance which is near to zero. With the same assumption, R_{b1}, R_{b2}, are affectless, because $(R_{b1}, R_{b2})>>1$. By considering mentioned assumptions the differential gain at the output of the LNA is obtained in (8).

$$Z_{in}=\frac{g_{m1}L_2}{C_{be1}}+\frac{(L_1+L_2)C_{be1}S^2+1}{SC_{be1}} \tag{7}$$

$$\frac{V_{out+}-V_{out-}}{V_{in}}=\left(\frac{g_{m1}\left(\frac{1}{SC_1}\,\|\,L_3S\,\|\left(\frac{1}{SC_{im}}+L_{im}S\right)\|\,Z_b\right)}{\left(1+g_{m1}L_2S+\left(L_1+L_2\right)C_{be1}S^2\right)}\times\frac{g_{m3}\left(Z_T\,\|\,\frac{1}{SC_4}\right)}{\left(1+g_{m3}L_4S+C_{be3}S^2\right)}\right)$$

$$Z_b=\frac{g_{m3}L_4}{C_{be3}}+\frac{L_4C_{be1}S^2+1}{SC_{be3}}$$

$$Z_T=\frac{S^2\left(L_5L_6-M^2\right)+SL_1R_L}{SL_6+R_L} \tag{8}$$

The variable transconductances $G_{M\text{-}I}$ and $G_{M\text{-}Q}$ amplifier (current combining stage) is shown in Fig. 7. The signals from amplifiers are collected and added through resistors R_o. As it can be proven, $G_{M\text{-}I}$ is obtained using equations (9). In this equation V_T expresses the thermal voltage which is about 26 mV at room temperature. The same formulation also can be manipulated for $G_{M\text{-}Q}$.

If ($V_d << 2V_T$), *tanh(x)*, can be approximated by its argument x, therefore in equation (9), $G_{M\text{-}I,Q1,2}=(I_{EE\text{-}I}/2V_T)$ and by substituting this result in equation (3), the desired phase shift can be adjusted just by ($I_{EE\text{-}Q}/I_{EE\text{-}I}$) which is concluded in (10). It should be mentioned (Q_1-Q_2) in conjunction with (Q_7-Q_8) are providing a positive phase direction while other transistors can produce a negative one.

$$G_{M-I,Q_{12}}=\frac{d\left(I_{c1}-I_{c2}\right)}{d\left(V_{inI+}-V_{inI-}\right)} \tag{9}$$

$$I_{c1}-I_{c2}=I_{EE-I}\times\tanh\left(\frac{\overbrace{V_{inI+}-V_{inI-}}^{V_d}}{2V_T}\right)$$

$$\theta_{out}=\pm a\,\mathrm{rctan}\left(\frac{I_{EE-Q}}{I_{EE-I}}\right) \tag{10}$$

III. Simulation Results

The proposed MCAN in Fig. 3(b), has been employed in a low-IF receiver to synthesize the desired phase shift as part of a vector modulator. The rest of circuits which is illustrated in Fig. 2 has been implemented at circuit level, in 0.13 µm SiGe-BiCMOS technology with seven metal layers and a supply voltage of 2 V. The circuit parameters such as the size of transistors, inductors, capacitors and resistors as well as the biasing voltages are optimized using SPECTRE simulations and equations (3)-(10), considering minimum phase error and a sufficient NF with the lowest value for power consumption. The differential gain versus frequency response of the receiver is indicated in Fig. 8. A voltage gain of 30 dB is obtained with a 3-dB bandwidth of 1.54 GHz at 28 GHz. This figure also reveals that a more than 40 dB rejection is also achieved at 22.275 GHz, which can be used for locating an image signal in the frequency plan of a low–IF receiver. The entire receiver consumes 8 mA from a 2 V supply voltage. Receiver's NF and input return loss (S_{11}) are demonstrated in Fig. 9 and Fig. 10, respectively. The receiver exhibits an NF less than 3.5 dB in the frequency range of 27-29 GHz. At 28 GHz the NF is lower than 2.8 dB, and the input return loss remains less than -17dB from 27-29GHz. The impact of a large amount of power at the input of the LNA on the phase error of the receiver is examined as illustrated in Fig. 11.

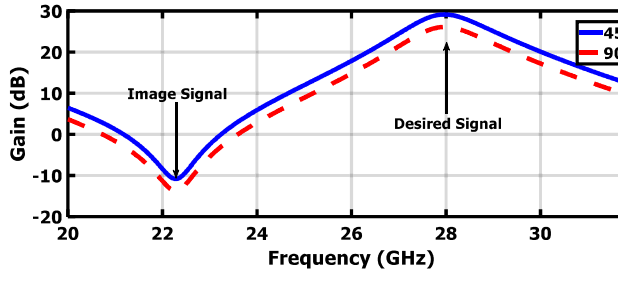


Fig. 8. Frequency response of the rest of the receiver chain for two different phase states.

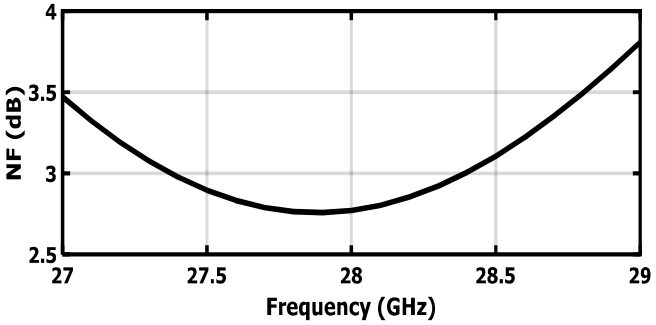


Fig. 9. NF of the receiver versus frequency. Achieving less than 2.8 dB at 28 GHz.

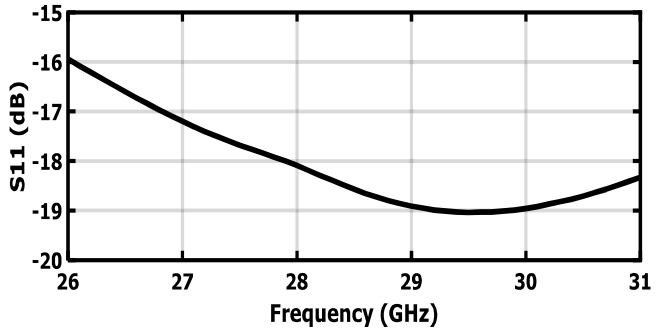


Fig. 10. Input return loss (S_{11}) versus frequency.

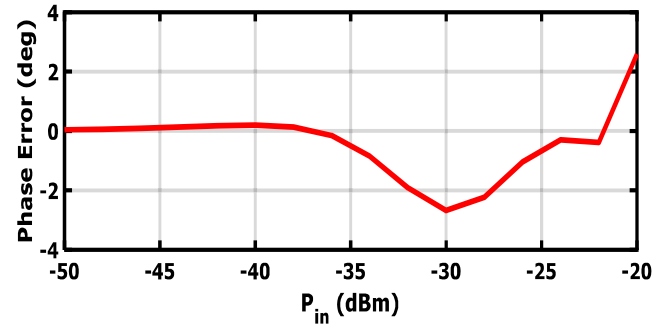


Fig. 11. Phase error at the output of the receiver versus input power.

The absolute phase error is about 2° which shows a relatively linear performance for cascaded stages, especially for transconductance stage. Fig. 12 depicts that the input-referred third-order intercept point (IIP3) of the receiver with a gain of 30 dB and around 28 GHz based on two-tone simulation with frequency spacing of 1 MHz is about -27.35 dBm. Performance of the receiver concerning image rejection is examined using discrete-time Fourier transform (DFT) which is indicated in Fig. 13. At the input of the receiver two sinusoidal tones have been applied with the same power (-30 dBm), one at 28 GHz and another at 22.275 GHz. A 512 points DFT with rectangular windowing is employed in the time period of 10 ns transient simulation. As it can be seen from Fig. 13, 40 dB rejection is observed at the image signal, which verifies the small-signal simulation.

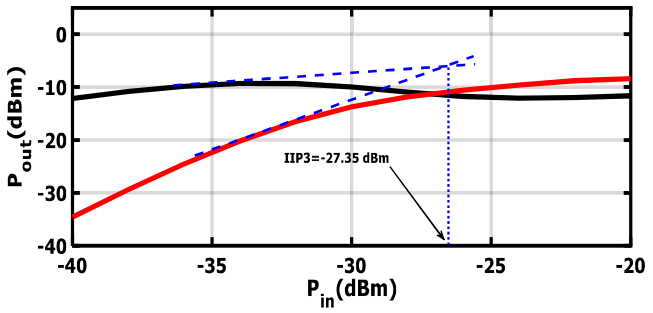


Fig. 12. IIP3 of the receiver around 28 GHz based on two-tone simulation with frequency spacing of 1 MHz.

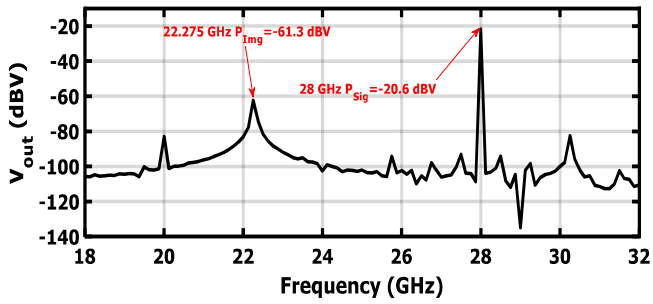


Fig. 13. DFT of the output of reciver voltage for two input sinusoidal tones with the same power (-30 dBm), one at 28 GHz and the other at 22.275 GHz.

IV. Conclusion

In this paper, a mutually coupled all-pass network (MCAN) has been presented to synthesize a quadrature IQ signal employed in a vector modulator. The proposed circuit employs a pair of mutual inductors to generate 90° phase shift which can lead to an equal signal gain for both I and Q paths. It is in contrast with PPF structure where the gain of the I path is always less than the gain of the Q path. The efficiency of the proposed topology has been verified by analytical methods and a circuit realization in SiGe-BiCMOS technology. Simulation results of the MCAN inside of a low-IF receiver represent a gain of 30 dB at 28 GHz with 3-dB bandwidth of 1.54 GHz. The NF of the receiver is less than 3.5 dB from 27-29 GHz and at 28 GHz the NF is lower than 2.8 dB. At the input of the receiver, a two stages transformer based LNA has been used which provides a S_{11} less than -17dB while 40 dB attenuation is obtained at 22.275 GHz. IIP3 of the receiver with a power consumption of 16 mW at 28 GHz is about -27dBm. The large-signal phase error of the receiver in presence of a large signal input tone is approximately 2°.

Acknoledgements

The authors acknowledge financial funding from the Austrian BMVIT via FFG in the project TRITON. They also thank Franz Dielacher from Infineon IFAT for access to the design environment and for chip fabrication.

References

[1] Chun-Nien Chen, *et al* "38-GHz CMOS Linearized Receiver With IM3 Suppression, P1 dB/IP3/RR3 Enhancements, and Mitigation of QAM Constellation Diagram Distortion in 5G MMW Systems", *IEEE Trans. Microw. Theory Tech*, Early Access, April 2020.

[2] Susnata Mondal, Jeyanandh Paramesh, "A Reconfigurable 28-/37-GHz MMSE-Adaptive Hybrid-Beamforming Receiver for Carrier Aggregation and Multi-Standard MIMO Communication", *IEEE J. Solid-State Circuits*, vol. 54, no. 5, pp. 1391–1406, May 2019.

[3] Domenico Pepe, Domenico Zito, "Two mm-Wave Vector Modulator Active Phase Shifters With Novel IQ Generator in 28 nm FDSOI CMOS", *IEEE J. Solid-State Circuits*, vol. 52, no. 2, pp. 344–356, Feb. 2017.

[4] Tianjun Wu, *et al* "A 51.5 - 64.5 GHz Active Phase Shifter Using Linear Phase Control Technique With 1.4° Phase resolution in 65-nm CMOS", IEEE RFIC, pp. 59-62, June 2019.

[5] Li Gao, Qian Ma, Gabriel M. Rebeiz, "A 20–44-GHz Image-Rejection Receiver With >75-dB Image-Rejection Ratio in 22-nm CMOS FD-SOI for 5G Applications", *IEEE Trans. Microw. Theory Tech*, Early Access, March 2020.

[6] Rahul Singh, Susnata Mondal , and Jeyanandh Paramesh, "A Compact Digitally-Assisted Merged LNA Vector Modulator Using Coupled Resonators for Integrated Beamforming Transceivers", *IEEE Trans. Microw. Theory Tech*, vol. 54, no. 5, pp. 2555–2567, July 2019.

[7] Kwang-Jin Koh, Gabriel M. Rebeiz, "0.13-um CMOS Phase Shifters for X-, Ku-, and K-Band Phased Arrays", *IEEE J. Solid-State Circuits*, vol. 42, no. 11, pp. 2535–2546, NOV. 2007.

[8] Timothy M. Hancock, and Gabriel M. Rebeiz, "A 12-GHz SiGe Phase Shifter With Integrated LNA", *IEEE Trans. Microw. Theory Tech*, vol. 53, no. 3, pp. 977–983, March 2005.

[9] Shailesh Kulkarni, *et al* "Design of an Optimal Layout Polyphase Filter for Millimeter-Wave Quadrature LO Generation" *IEEE Trans. Circuits Syst. II,* vol. 60, no. 4, pp. 202–206, April. 2013.

Flow-Aware QoS Engine for Ultra-Dense SDN Scenarios

Mertkan AKKOÇ and Berk CANBERK
Computer Engineering Department
Faculty of Computer and Informatics
Istanbul Technical University
Ayazaga 34469, Istanbul-Turkey
Email: {akkocm, canberk}@itu.edu.tr

***Abstract*—Software-Defined Networking (SDN) is a promising technology for 5G thanks to the separation of data plane and control plane. However, especially in ultra-dense scenarios, as a result of the centrality of the SDN controller, the response time increases with the ultra high spiky demands of the incoming heterogeneous flows. This sudden increase causes an uncontrollable rise in end-to-end (e2e) latency and drop rate in the controller. Moreover, this also leads to an unbalanced QoS provisioning in this heterogeneous flow management. To tackle these challenges, in this paper, we propose a Flow-Aware QoS Engine by considering both huge flow demands and QoS requirements of different 5G flows (URLLC, eMBB, mMTC). This novel QoS-based engine contains two steps in a single controller: The Admission Management and The Priority Management. In admission management, we modify the generic Loss Ratio-Based Random Early Detection Algorithm (LRED) by implementing an add-on containing the arrival rate of different flow types. In the proposed priority management, we design a tree-based priority management scheme where we dynamically assign priorities to each flow type regarding fairness. According to our thorough evaluation results, we get up to 53% lower response times, up to 58% lower e2e latencies, and up to 36% lower drop rates for three different flows in ultra-dense SDN scenarios.**

***Index Terms*—5G, software defined networking (SDN), controller response time, admission management, priority management, fairness**

I. Introduction

5G is a promising technology that allows 1~20 Gbps throughput and ultra-low latency, which is less than 1 ms [1]. It is expected from 5G that it can provide tailored resources to meet the QoS requirements of different services, specifically in ultra-dense scenarios. Here, heterogeneous flows have been defined: Enhanced Mobile Broad (eMBB) that needs 4ms latency for video/audio flow; Massive Machine-Type Communication (mMTC) that needs 10 ms latency for flow contains massive but small packets; Ultra-Reliable and Low-Latency Communications (URLLC) that requires 0.5 ms latency for remote control flow of drones, robots, factories, etc. [2]. To meet these requirements, one of the technologies that 5G used is Software Defined Networking. But, when there is huge flow density in the network, the SDN controller can give a slower response to incoming packets in the control plane as a result of having a global view, and centrality.

In the ultra-dense areas, the number of flows in the network quickly and uncontrollably increases because of ultra-high demand in Radio Access Network (RAN). This high-demand can cause an increase in the number of packet_in messages in the SDN. Because a flow entry that can match with the incoming flow could be deleted from the OpenFlow table, or some flows can actually be new in the network. Path computation algorithms in the SDN controller can't adapt themselves against this increase. As a result, the controller doesn't meet the requirements of different 5G flows, and the response time of the controller decreases [3]. Further, congestion may occur in the SDN controller, and the network becomes less robust, elastic, and unfair. As a result, the drop rate of packets increases, which means users can experience less data rate and higher e2e latency. To see an increase in response time of the SDN controller, we create a network described in Section III, but we did not implement any solution method for the interested problem. As seen in Fig.1, if we increase the number of hosts i.e., flow requests, and it is greater than 200, we exceed the target response time (10 ms)[1].

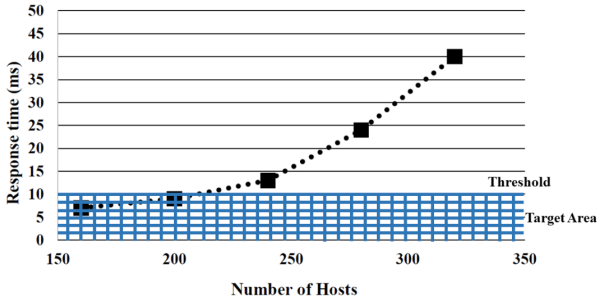


Fig. 1: SDN Controller Response Time vs Host Number

Thus, through this paper, we aim to manage the processing of the control plane faster for newly incoming packets in ultra-dense SDN scenarios.

[1]For the controller response time, 10 ms is selected based on the latency requirement of mMTC flow type which is the highest requirement among different flow types.

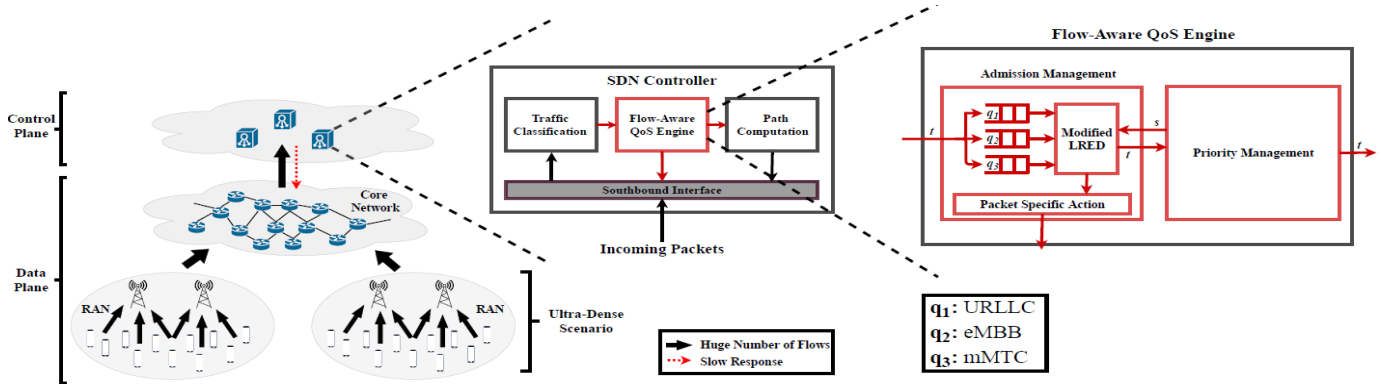


Fig. 2: The proposed system model

A. Related Work

As mentioned, one of the main reasons for the increase in e2e latency and drop rate of packets when there is a huge number of flows in the network is an increase in response time of the control plane. To handle more packets in the control plane, the most accepted approach is distributed SDN controller system architecture [4]. In this concept, [4] tries to balance the load, considering the response time of controllers as a threshold value and proposes a solution for the overloaded situations. In [5], wild-card entries are installed to switches for flows, but not all flows before they arrived to the switch to decrease the load of the controller. [6] proposes a rounding-based algorithm to equalize the load in links and also in controllers. In [7], there is a two-state solution. First, they solve the stable matching problem for switches and controllers; then, they try to reach Nash stable point by changing switch-controller assignment. [8] proposes a new metric named Quality of Controller (QoC) depends on both reliability and response time of controller and doing switch-controller mapping considering QoC. [9] matches switches with controllers in a manner of one-to-many matching and takes the processing capacity of controllers as a criterion. [10] predicts loads of switches, and among controllers, they migrate switches, which possibly cause congestion in controllers assigned these switches according to predictions. In [11] in which an intermediate layer named "flowcache" is proposed aims to store last incoming flows here, thereby reducing the capacity of the controller. Also, [12] proposes an architecture where there are different types of controllers named "Area Controller", "Supervised Controller", and "Contributory Controller" in. However, none of these works consider the requirements of QoS of three flow types defined for 5G services. Also, solutions of these works are more complex; as a result, slower than our proposed solution. Lastly, they don't prioritize flows, both considering delay tolerance and future states of buffers (queues) of different flow types.

B. Contributions

In this paper, we focus on decreasing the packet drop rate and e2e latency by reducing the response time of the SDN controller in ultra-dense areas, which means when the number of newly arrived flows uncontrollably increase. To achieve these aims, we create a flow-aware QoS engine implemented in the SDN controller that contains two steps: admission and prioritization. In admission, we use Loss Ratio-based RED (LRED) queue management in the [13]. In prioritization, we construct a tree that contains states of the near future of buffers in this step and prioritize packets of each flow type considering these buffer states and priority values of flow types. So, our contributions include:

- In the admission step, we evaluate the LRED for multiple queues to make the network more robust and elastic. Also, we take into account states of queues in the priority management. Here, we prevent high response time by decreasing the congestion in the SDN controller.
- We define a fairness value, which considers the delay tolerance of the different flow types of 5G. Based on this value, we create novel priority management to make the network fairer to different flow types. This fair flow assessment leads to a lower drop rate and e2e latency.

We will investigate the proposed flow-aware QoS engine in detail in Section II. In Section III, the results of the performance analysis will be examined. In Section IV, we will conclude the paper.

II. Flow-Aware QoS Engine

The proposed flow-aware QoS engine consists of two modules: admission and priority management, as seen in Fig.2. In this engine, we take into account of requirements of flows of eMBB, mMTC, URLLC services. We first put an admission module implemented with a modified LRED algorithm to prevent congestion. In this module, we also consider the packet loss ratio of each flow type and queue length of each flow type in the priority management module (s) while calculating the drop probability of each type. When this module decides that a packet of some flow type has to drop, it sends packet-specific action to the switch. If not, it sends a packet (t) to the priority management module. In the priority management module, we take into account the delay of each flow type in the queue while deciding priority. In this module, we also consider possible states of queues in the near future for priority.

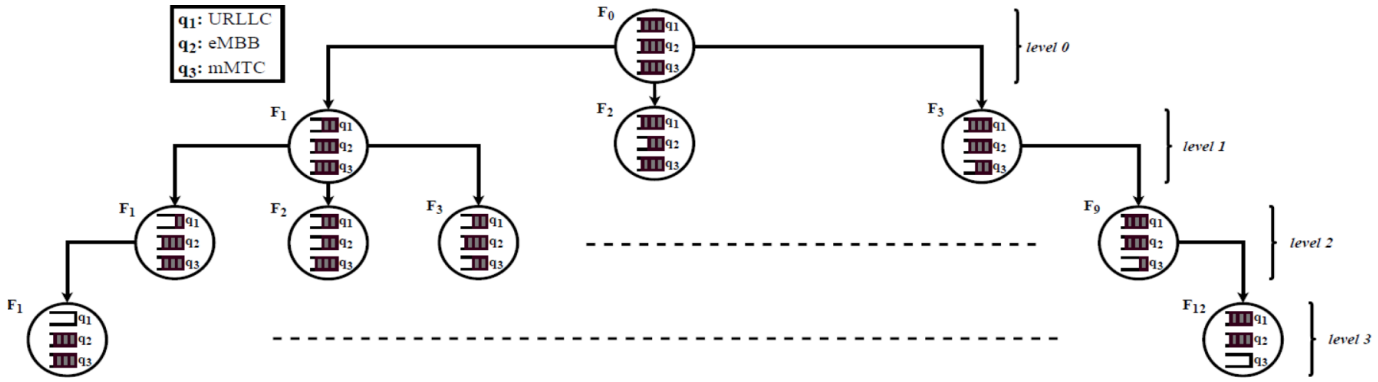


Fig. 3: Tree structure for priority management

A. Admission Management

For the acceptance of incoming packets, we modified the LRED algorithm, which is one of the active queue management algorithms, according to the multi-queue status. We constitute these queues for the flows of URLLC, eMBB, and mMTC services separately. The proposed algorithm can be seen in Alg. 1. In the algorithm, after calculating the drop probability foreach flow type, we check whether each queue is full or not. (By saying drop probability, we mean that the SDN controller just ignores packet_in message for the newly arrived flow. After this point, we use drop and ignore interchangeably.) If a queue is full, we ignore the packet_in message without looking at any other condition; else, we check whether there is a signal that comes from priority management. If there is a signal represented with a Boolean value s in Alg.1, that means queue of this flow type in the priority management module is full; we again ignore the packet_in message to prevent a bottleneck in priority management step. Finally, we compare a random value, produced uniformly, and is between 0 and 1, with the drop probability. If this random value is less than the drop probability, we ignore the message. We increase the number of ignored messages one by one when we ignore a message because we will use this number in the calculation of the drop ratio. We maybe cause to unnecessarily drop whole packets of a flow, which is one of the other flows come to a switch by ignoring the packet_in message of this flow whether the queue in the controller is full or not. But if we ignore the packet_in message only when the queue of this flow type is full, we can cause to drop all flows that come to that switch.

To calculate the drop probability of each flow type in Alg. 1, we use:

$$p_i = m_i^*(t) + \alpha\sqrt{m_i^*(t)}(q_i - q_{wi}), \quad \alpha > 0 \tag{1}$$

In the Eq. 1, α is a constant number. q_i represents the instant queue length; q_{wi} is the queue length that we wanted to be in steady-state for each flow type. Unlike [13], we use different q_{wi} values for each flow types. Because, arrival rate of each flow type may be different from each other. So, we use the lowest q_{wi} value for the flow type that has the highest arrival rate. As a result, we prevent unnecessary packet drops for the flow type that has the highest arrival rate. Finally, $m_i^*(t)$ means that the expected drop rate in the t^{th} period for each flow type. We calculate this value using the equation:

$$m_i^*(t) = w_i.m_i^*(t-1) + (1 - wi).m_i(t) \tag{2}$$

In the Eq. 2, w_i is the weighting factor and $m_i(t)$ is the drop ratio for each flow type, that is calculated after latest O period and calculated using the equation:

$$m_i(t) = \frac{\sum_{n=0}^{O-1} l_{di}(t-n)}{\sum_{n=0}^{O-1} l_{ai}(t-n)} \tag{3}$$

In the Eq. 3, while $l_{di}(t)$ is the number of ignored packet_in messages of each flow type, $l_{ai}(t)$ is the number of all packet_in messages of each flow type that come to the SDN controller.

Algorithm 1 Admission Management Algorithm

```
Initialize α
set s is FALSE
for each flow type i do
    calculate the drop probability p_i using equation (1)
    produce a random value r between 0 and 1 uniformly
    if queue is not empty then
        ignore the packet_in message
        increase l_di by one
    else
        if s_i is TRUE then
            ignore the packet_in message
            increase l_di by one
        if r < p_i then
            ignore the packet_in message
            increase l_di by one
end
```

B. Priority Management

In this module, we aim to achieve two main objectives: behaving each flow type according to their delay tolerance and decreasing delay of all packets in the SDN controller. We created three different queues for each different flow types to achieve the former main objective. And also, we gave different priority values to these queues considering delay tolerance of each flow. But creating queues brings about the problem of

keeping each queue length below a certain level. We also solve this problem while realizing the latter main objective.

We constructed a tree structure that contains three levels except for the starting level, like in Fig.5. If there were more levels in the tree, the algorithm of this method had to calculate more possible numbers and remember these numbers. As a result, we would have a slower method. To prevent this, we decide that three is the optimum number of levels. Each level represents future steps and has three possible results derived from different states of the queue. Because these results also represent states of the queue, we can say three possible states instead of results. At the first level or first future step, there are three possible states: only picking a packet from the first queue or only from the second queue or only from the third queue. As seen from the Fig.3, we assume that there is no incoming packet for all queues in the future states. We show a value in each node that we put the name of "fairness value (F)". We obtain this value using:

$$F = max(N_1 o_1, N_2 o_2, N_3 o_3) max(\frac{N_1}{o_1}, \frac{N_2}{o_2}, \frac{N_3}{o_3}) \quad (4)$$

In the Eq. 4, N means queue length; o holds the priority value assigned to queues considering the delay tolerance of flow types. By using queue lengths, we take into account the delay of packets in the priority management module. The priority calculation is biased to 1, and each priority value (o) is between 0 and 1; that means $\sum_{n=1}^{3} o_n = 1$. While $max(.)$ function at the right cause to choose a packet from the queue that has the highest priority value; $max(.)$ function at the left cause to choose a packet from queue that has the highest length. Because we choose the node that has the highest fairness value from a level as a parent node to generate possible states of the next level. Thus, we provide fairness among flow types considering both queue lengths and priority values of them.

Algorithm 2 Priority Management Algorithm

get N_1^0, N_2^0, N_3^0 for each queue
for l = 1 to 3 **do**
  calculate $N_{possible}^l$ for each queue in each node
  if there are more than one node with $N_{possible}^l \geq N_{threshold}$ **then**
    choose the node that has more queues with higher priority value as a parent
    set decision path
  if there is one node has queue(s) with $N_{possible}^l \geq N_{threshold}$ **then**
    choose that node as a parent node
    set decision path
  else
    calculate $F_{possible}^l$
    choose the node that has max F as a parent node
    set decision path
do decision path
end

We try to achieve the second main objective by using queue lengths, priority values, and constructing a tree. Thanks to the tree, we foresee all possible future states of queues. Then, we choose the best possible path, which means the best next three moves, from the start point to the lowest level. After selecting these moves, we do not calculate the possible fairness values again during the process of realizing these three moves. But calculating all possible fairness values and traversing the tree to choose the best path has a decreasing effect on the delay of packets. To prevent this adverse effect, we create the 'Priority Management Algorithm' in the Alg.2. In this algorithm, there are three ways to choose a node as a parent node to calculate next possible states: *i-)*if there are more than one nodes that have queue(s) with a queue length ($N_{possible}^l$) which is greater than a threshold value ($N_{threshold}$), we choose the node that has more queues with a higher priority value. We determine the threshold value is 9(*maximum queue length*)/10. Because we expect a decrease in the length of all queues until this threshold value. The reason for this expectation is the ratio of the possible highest priority value (0.7) with the lowest one (0.1) assuming each flow type has a different priority value. *ii-)* if there is one node in this state, we choose that node. *iii-)* if there is no node in this state, we choose the node with has the highest fairness value as a parent node.

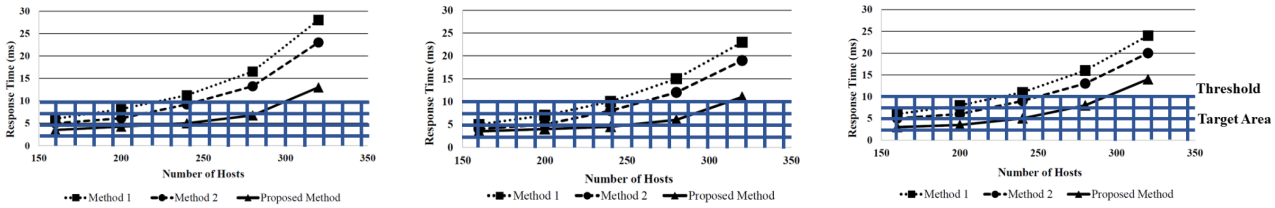
III. Performance Evaluation

We implement a network with the topology that there are 20 switches in the data plane and 2 controllers in the control plane to show that we can also use the proposed method in distributed SDN control architecture. 10 switches are assigned to each controller. Each switch has 8 hosts, and each host sends 4 flow. We implement this network on a single machine that has Ubuntu 18.04 LTS, Intel Core i7, 12.00 GB RAM. We use *POX* [14] as a controller. We created three flow types: URLLC, mMTC, and eMBB in the network. As a priority value, we give 0.5 to URLLC; 0.3 to eMBB and 0.2 to mMTC flow. We set the lengths of each queue to 1000 packets and adjust the threshold value in the priority management, as mentioned in Section II. Details of other simulation parameters are given in Table 1. We choose different packet sizes for different network services to reflect the characteristics of these services to the performance evaluation. The reason of giving these priority values and creating packets with these sizes is characteristics and requirements of flow types mentioned in the Section I. Since URLLC and mMTc flows are TCP, but eMBB is UDP flow, we generated these flows with different distributions according to "waiting time calculation" in [15].

TABLE I: Simulation Parameters

Mininet Version	2.2.2		
OpenFlow Version	1.3		
Hosts per Switch	[8 – 16]		
λ per Flow	500 packet/sec		
Packet Sizes	URLLC	mMTC	eMBB
	100 bytes	80 bytes	70 bytes

To compare our solution in terms of controller response time to each flow type, e2e latency and drop rate of packets of each flow type, we implement three different methods: implementing First In First Out (FIFO) model in admission module without priority module (method 1); implementing our modified LRED model in admission module without priority



(a) URLLC (b) eMBB (c) mMTC

Fig. 4: Average Response Time of the Controllers to Different flow Types vs. Number of Hosts

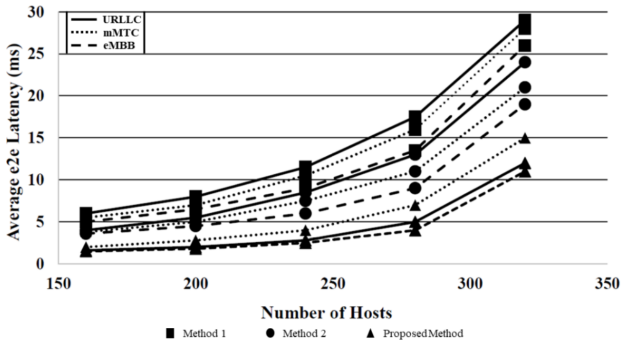


Fig. 5: e2e Latency vs Number of Hosts

module (method 2), and implementing proposed flow-aware QoS engine (proposed method). We increase the number of hosts adding 2 hosts to each switch for each experiment, like in [12]. As seen in Fig. 4, the average response time of the SDN controllers in the proposed method than other methods. The reason for getting the best result and very close to the target area, described in Section I, is that we don't admit packets if buffers of all flow types are full in the priority management. Also, we give the highest priority to URLLC flow in the Eq.4. As a result, we get the best improvement in response time of the SDN controller for URLLC flow (approximately 43 % better than method 2, 53% better than method 1 for 3200 hosts). Similar improvements can be seen for other flow types in Fig.4b and Fig.4c. The reason for not getting huge differences between results for flow types that means fair results for them is the ratio between queue length and priority value in Eq.4.

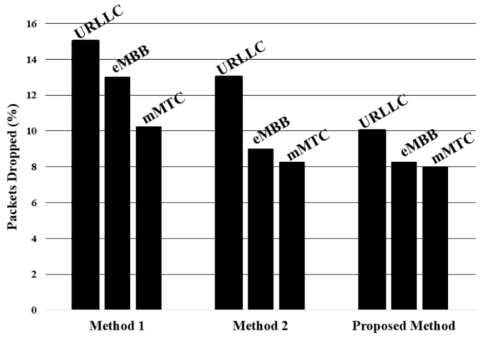


Fig. 6: Packet Drop Rate

As seen in Fig.5, we get the best result in e2e latency for all flow types. Because we take into account the packet drop rate of each flow type using the Eq.3 and calculate different packet drop probability for each flow type using the Eq.1. As a result, we respond faster and more adaptable to the threat of congestion in the controller. Here, we get the best enhancement for URLLC flow (approximately 50% better than method 2, 58% better than method 1 for 3200 hosts), and the worst enhancement for mMTC flow (approximately 28% better than method 2, 46% better than method 1 for 3200 hosts) because of priority values used in the Eq.4. Similar improvements can be seen for eMBB flow in Fig.5. Also, in the admission management module, we consider the states of queues of each flow type in the priority management to decide whether we drop the packet or not in the Alg. 1. But, these considerations give a disadvantage in the packet drop rate although we got the best result. As seen in Fig.6, we get much better results for all flow types in the proposed method than method 1, but not so much better results than method 2. For example, we get approximately 21% better result in the proposed method than method 1; but approximately 3% better result than method 2 for mMTC flow. Similar improvements can be seen for other flow types in Fig.6.

IV. Conclusion

In this paper, we propose a fair and rapid QoS provisioning solution for the SDN controllers facing heterogeneous service flows in ultra-dense scenarios. We develop a novel flow-aware QoS Engine to prevent congestions in the controller when heterogeneous URLLC, eMBB, and mMTC traffic suddenly increase. The fair processing of the proposed engine for heterogeneous flows provides faster response time to the incoming new packets (up to 53%). Also, as a result of faster response time, we decrease the e2e latency (up to 58%) and drop rates (up to 36%).

Acknowledgment

This work was supported by ITU Scientific Research Fund with a project number: 42439. Also, at this work, Mertkan Akkoç was supported by the Turkcell-Istanbul Technical University Researcher Funding Program.

References

[1] Cisco ultra 5g packet core solution white paper. Technical report, Cisco, 2018.

[2] S. Lien, S. Hung, D. Deng, and Y. J. Wang. Efficient ultra-reliable and low latency communications and massive machine-type communications in 5g new radio. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–7, Dec 2017.

[3] Paul Göransson, Chuck Black, and Timothy Culver. Chapter 12 - sdn applications. In Paul Göransson, Chuck Black, and Timothy Culver, editors, *Software Defined Networks (Second Edition)*, pages 271 – 301. Morgan Kaufmann, Boston, second edition edition, 2017.

[4] J. Cui, Q. Lu, H. Zhong, M. Tian, and L. Liu. A load-balancing mechanism for distributed sdn control plane using response time. *IEEE Transactions on Network and Service Management*, 15(4):1197–1206, Dec 2018.

[5] Hongli Xu, Jianchun Liu, Chen Qian, He Huang, and Chunming Qiao. Reducing controller response time with hybrid routing in software defined networks. *Computer Networks*, 164:106891, 2019.

[6] Haibo Wang, Hongli Xu, Liusheng Huang, Jianxin Wang, and Xuwei Yang. Load-balancing routing in software defined networks with multiple controllers. *Computer Networks*, 141:82 – 91, 2018.

[7] T. Wang, F. Liu, and H. Xu. An efficient online algorithm for dynamic sdn controller assignment in data center networks. *IEEE/ACM Transactions on Networking*, 25(5):2788–2801, Oct 2017.

[8] V. Sridharan, P. M. Mohan, and M. Gurusamy. Qoc-aware control traffic engineering in software defined networks. *IEEE Transactions on Network and Service Management*, pages 1–1, 2019.

[9] A. Filali, A. Kobbane, M. Elmachkour, and S. Cherkaoui. Sdn controller assignment and load balancing with minimum quota of processing capacity. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2018.

[10] A. Filali, S. Cherkaoui, and A. Kobbane. Prediction-based switch migration scheduling for sdn load balancing. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2019.

[11] A. Ruia, C. J. Casey, S. Saha, and A. Sprintson. Flowcache: A cache-based approach for improving sdn scalability. In *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 610–615, April 2016.

[12] Amit Nayyer, Aman Kumar Sharma, and Lalit Kumar Awasthi. Laman: A supervisor controller based scalable framework for software defined networks. *Computer Networks*, 159:125 – 134, 2019.

[13] C. Wang, J. Liu, B. Li, K. Sohraby, and Y. T. Hou. Lred: A robust and responsive aqm algorithm using packet loss ratio measurement. *IEEE Transactions on Parallel and Distributed Systems*, 18(1):29–43, Jan 2007.

[14] The pox network software platform. https://noxrepo.github.io/pox-doc/html/.

[15] Müge Erel Özçevik, Berk Canberk, and Trung Q. Duong. End to end delay modeling of heterogeneous traffic flows in software defined 5g networks. *Ad Hoc Networks*, 60:26 – 39, 2017.

Network Bandwidth Usage Forecast in Content Delivery Networks

Aykut Teker*, Ahmet Haydar Örnek*, Berk Canberk*†
*Medianova CDN R&D Center, Şehit Ahmet Sokak No:4 Mecidiyeköy İş Mer. 15th Floor, Şişli/İstanbul/Turkey
†Faculty of Computer and Informatics Engineering, Istanbul Technical University, Istanbul/Turkey
Email: aykut.teker@medianova.com, ahmet.ornek@medianova.com, canberk@itu.edu.tr

***Abstract*—Operational burden of a Content Delivery Network that is a vast overlay network on top of current Internet Architecture can be alleviated by forecasting Content Delivery Network bandwidths. The purpose of this paper is to forecast network bandwidth usage for Content Delivery Networks' Points of Presence. In this paper we compare Seasonal Auto-Regressive Integrated Moving Averages and Artificial Neural Networks that can be used for predicting and minimizing operational costs of Content Delivery Networks via resource allocation, server allotment and local ISP bandwidth contract costs. We directly forecast end-user to Content Delivery Network bandwidth, so it can directly be used to lower end-user latencies. In this paper; we first conduct Self-Similarity Analysis and then utilize Seasonal Auto-Regressive Integrated Moving Averages and Artificial Neural Networks to predict bandwidth usage with 6.338% error.**

***Index Terms*—content delivery networks, artificial neural networks, seasonal auto-regressive integrated moving averages, traffic modelling.**

I. INTRODUCTION

Content providers lease their proprietary material to Content Delivery Networks (CDNs) within the agreed-upon delivery time limits that are laid out in Service Level Agreements (SLA). To meet contractual obligations, CDNs use an overlay network on top of the Internet Architecture that employs globally widespread Points of Presence (PoP). A sketch of CDN architecture is presented in Fig. 1 –where different types of PoPs are shown as Origin, Mid-Cache and Edge. PoPs at different locations are subject to local constrictions on internet speed by their corresponding Internet Service Providers (ISPs) and variances in internet speed directly affects the overall performance of CDNs which can directly be monitored by longer end-user latencies. Such differences in traffic speed also reflects to CDN operators as extra costs, since CDN PoPs work sub-optimally without traffic forecast. In Fig. 2; we present 3-month out-traffic from a PoP within Medianova CDN; red dashed-line is ISP-imposed speed cap that reflects as a direct cost surplus.

According to the challenges mentioned above, our contributions can be summarized as follows:

- We forecast using CDN PoP traffic data between the end-user and a PoP; therefore our PoP bandwidth forecast results are more dependable since it is directly in between the end-user and the PoP.
- Our approach is more reliable as it also tacitly incorporates local ISP-imposed speed restrictions on the traffic. We do this by using out-traffic data from the PoP; hence only content delivery is subject to ISP traffic speeds.
- Our forecast results can directly be used to reduce end-user latencies by redirecting traffic to PoPs with faster internet connection.

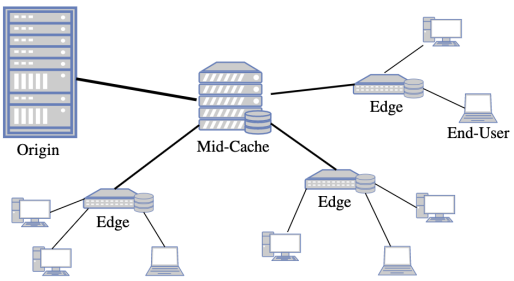


Fig. 1: A sketch of Content Delivery Networks as an overlay network that relay content to end-users via different Points of Presence; *i.e.* Origin, Mid-Cache, and Edge servers.

CDN traffic forecast can lead to more optimal CDN operation: CDNs can use forecast methodology to reduce their local ISP lease contracts; future PoP bandwidth usages can be accurately determined via numerical methods calculated in this paper. Furthermore; CDN traffic forecast can be used to design a more efficient CDN architecture implemented by means of resource allocation, request routing, and server allotment.

This paper is organized as follows; in Section II we summarize current literature about forecasting techniques mentioning their diverse use cases. In Section III, we lay out our CDN bandwidth forecast methodology; present our results using two different forecasting methods. Section IV is dedicated to analyzing our results and pondering on the most accurate method for different forecast horizons. Finally, in Section V, we conclude our results.

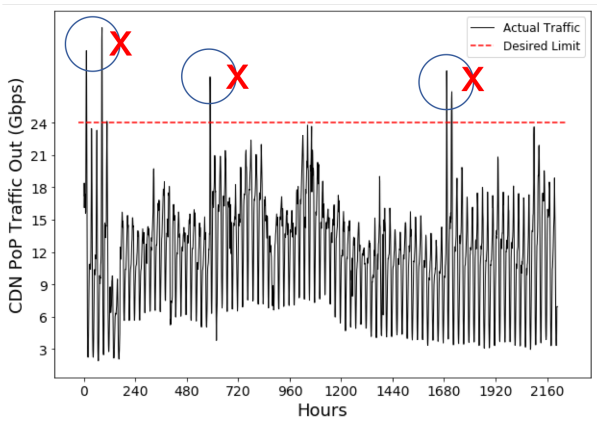


Fig. 2: Three-month out-traffic bandwidth (Gbps) usage of a PoP. Red dashed-line is the bandwidth cap that is imposed by the local Internet Service Provider. Excess usage of CDN PoP bandwidth above the cap leads to more operational cost.

II. RELATED WORK

Traffic forecast has been studied in the literature; however, it has never been studied using actual data from CDN PoPs. Predicting CDN traffic is practically univariate Time-Series Forecasting (TSF) as it constitutes of analysing N-pieces of traffic data in order to predict the $N+1^{\text{th}}$ time-series traffic. In literature; Time-Series Forecasting (TSF) has been employed for various types of predictions ranging from network security to econometrics; however we will first mention network-oriented approaches.

Different modeling tools have been used to forecast different types of internet traffic. In [1], they employ Auto-Regressive Integrated Moving Averages (ARIMA) and Artificial Neural Networks (ANN) to capture cloud computing traffic behavior. [2] and [3] use various types of Moving-Averages (MA) also to forecast cloud computing network bandwidth. Computer network traffic has been modeled in [4]–[6] using Discrete Wavelet Transform (DWT), Recurrent Neural Networks (RNN) and Auto-Regressive Moving Averages (ARMA) for computer networks, and in [7] using Generalized Cauchy Processes (GCP). Apart from wired networks; [8] presents short-term traffic prediction for wireless networks using different MA formulations.

Modeling formulations such as ARMA, ARIMA, MA are dubbed as *traditional approaches* in TSF; whereas RNN and ANN methods are more current approaches. Internet traffic is predicted via TCP/IP protocol using neural networks in [9] and reports that ANNs are as compatible as ARIMA methods in TSF. [10] and [11] have implemented ANNs to forecast nonlinear time-series by varying sample size, different number of input and hidden nodes, and they report that a larger validation set that is used to train the ANN overcomes the overfitting problem.

Within a CDN perspective the self-similar (SS) analysis of the CDN data becomes crucial as it has been incorporated into network modeling structure in [2], [3], [7], [8]. Self-similarity analysis becomes important as ARIMA and MA are *linear forecasting methods*; and applying such methods to nonlinear CDN PoP bandwidth data as in Fig. 2, leads to more accurate results if the network traffic is statistically analyzed. On the contrary, statistical analysis becomes redundant when forecasting with newer approaches such as ANN and RNN; nonlinear constructs they are, they are sufficient in *learning* from nonlinear PoP data.

We also want to make use of the periodic nature of the CDN PoP traffic as its fluctuating behavior can be seen in Fig. 2. That's why we shall use *Seasonal*-ARIMA (SARIMA) method, instead of ARIMA in our traditional approach. SARIMA method presents more accurate results with periodicity. CDN PoP bandwidth forecast with SARIMA method has never been utilized in literature; however SARIMA has been used for a variety of forecasts ranging from border inspection [12], cassava export [13] to wind-speed [14].

In the next section we first statistically analyze CDN PoP bandwidth data and then continue to present our forecast results using SARIMA and ANNs.

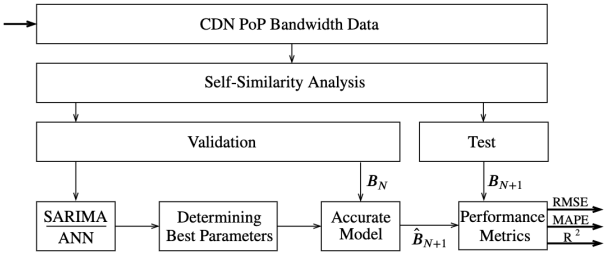


Fig. 3: A flowchart of our CDN PoP forecast methodology. PoP bandwidth data is first analyzed statistically and then split into Validation and Test sets. Validation set is used to fit SARIMA function by means of hyperparameters and the same is also used to train ANN framework. After finding the most accurate model; a forecast using data from the Validation set B_N is used to forecast $\hat{B}_{N+1}$ and via comparison an error analysis is made using Test set bandwidths B_{N+1} via Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and R^2.

III. METHODS AND RESULTS

In this paper, we use two bandwidth datasets taken from two PoPs within Medianova architecture. Our method is schematized in Fig. 3. We start by determining whether CDN data is self-similar or not; then we divide CDN data into *Validation* and *Test*. Validation set will be used to *fit* SARIMA function and *train* ANN. Specifically, SARIMA is fitted by means of Walk-Forward Validation which determines the most accurate SARIMA hyperparameters; and ANN is trained using a Rolling Window Approach per same premise. The model is then used to forecast with the most accurate models and Test set is used for error analysis. Before we present our results, we first mention performance metrics with which we determine the forecast errors.

A. Performance Metrics

We use three distinct performance metrics. Root Mean Squared Error (RMSE) *i.e.* $E_{\text{RMSE}} = (n^{-1}\sum_{m=N}(B_m - \hat{B}_m)^2)^{1/2}$ where B_m and $\hat{B}_m$ are actual and predicted bandwidths; respectively and N is the number of time-series data points in the Test set. Next; we use Mean Absolute Percentage Error (MAPE) as a performance metric *i.e.* $E_{\text{MAPE}} = (n^{-1}\sum_{m=N}|B_m - \hat{B}_m|/B_m) * 100$. n is the normalization factor in RMSE and MAPE that depends on the number of data points in the forecast horizon. Then, we also calculate R^2 scores for both validation and test sets as R^2 score is an important performance metric for ANNs. R^2 score lies within range $0 \leq R^2 \leq 1$, where results closer to unity represent more accurate predictions.

B. Self-Similarity Analysis of CDN Traffic

We start by analyzing the CDN PoP bandwidth data by means of self-similarity. We found literature as bereft of CDN self-similarity analysis, so we start by investigating whether CDN traffic is SS or not. Self-similarity analysis can be done by calculating the Hurst parameter (H) which is a measure of self-similarity in a network where H is a measure of the speed of decay of the Auto-Correlation Function (ACF). If

	Validation			Test		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
2-Day Forecast						
SARIMA$(\mathbf{2,0,2})(\mathbf{2,0,1})_{\mathbf{24}}$	191.87	2.9871	0.942	1035.2	7.213	0.955
SARIMA$(2,0,1)(2,0,2)_{24}$	184.98	3.0015	0.946	1190.9	7.743	0.940
SARIMA$(2,0,1)(2,1,1)_{24}$	196.14	3.0225	0.940	1093.1	8.724	0.950
SARIMA$(1,0,2)(1,0,2)_{24}$	185.71	3.0436	0.946	1037.5	7.250	0.955
8-Day Forecast						
SARIMA$(2,1,2)(1,0,1)_{24}$	77.688	1.6719	0.992	2500.8	16.678	0.760
SARIMA$(\mathbf{2,1,2})(\mathbf{2,1,1})_{\mathbf{24}}$	101.45	1.9009	0.987	2106.3	11.710	0.830
SARIMA$(1,1,2)(1,0,2)_{24}$	86.180	1.9238	0.991	2225.6	17.005	0.810
SARIMA$(2,1,1)(2,1,1)_{24}$	101.28	1.9477	0.987	2110.6	11.696	0.829
14-Day Forecast						
SARIMA$(1,0,2)(2,1,2)_{24}$	86.717	1.2714	0.991	1672.2	9.188	0.880
SARIMA$(\mathbf{1,0,2})(\mathbf{1,1,1})_{\mathbf{24}}$	85.581	1.2854	0.991	1650.2	9.023	0.883
SARIMA$(1,0,2)(2,1,1)_{24}$	87.772	1.2895	0.991	1681.0	9.323	0.879
SARIMA$(2,1,2)(2,1,1)_{24}$	59.892	1.2971	0.996	2231.0	11.857	0.786

TABLE I: For the first PoP; most accurate forecasts for 2-Day, 8-Day, and 14-Day forecast horizons obtained via SARIMA hyperparameter grid-search using Walk-Forward Validation. Inspecting Auto-Correlation and Partial Auto-Correlation Function plots, the hyperparameter grid search is done with $p/P, d/D, q/Q = \{1,2\},\{0,1\},\{1,2\}$ for $s = 24$.

	Validation			Test		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
2-Day Forecast						
ANN$(n_\mathrm{i}, n_\mathrm{o})$	1710.4	16.673	0.879	1732.1	16.629	0.878
ANN$(n_\mathrm{i}, 16, n_\mathrm{o})$	1098.8	8.7292	0.951	1357.4	10.947	0.925
ANN$(n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{o})$	1033.6	7.9893	0.955	1209.8	9.469	0.940
ANN$(\mathbf{n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{i}/2, n_\mathrm{o}})$	1088.2	7.6988	0.951	1222.7	9.320	0.939
8-Day Forecast						
ANN$(n_\mathrm{i}, n_\mathrm{o})$	1862.6	16.832	0.867	1950.1	17.547	0.855
ANN$(n_\mathrm{i}, 16, n_\mathrm{o})$	1920.4	12.584	0.858	2041.9	11.956	0.841
ANN$(\mathbf{n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{o}})$	1898.3	10.028	0.862	1888.9	11.324	0.863
ANN$(n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{i}/2, n_\mathrm{o})$	2074.1	10.928	0.835	2027.1	11.433	0.843
14-Day Forecast						
ANN$(n_\mathrm{i}, n_\mathrm{o})$	1951.5	17.994	0.837	1786.7	17.459	0.863
ANN$(n_\mathrm{i}, 16, n_\mathrm{o})$	1607.7	11.921	0.889	1535.1	10.085	0.899
ANN$(n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{o})$	1472.1	9.9998	0.907	1616.3	11.966	0.888
ANN$(\mathbf{n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{i}/2, n_\mathrm{o}})$	1192.2	7.6583	0.938	1384.1	8.308	0.918

TABLE II: For the first PoP; most accurate forecasts for 2-Day, 8-Day, and 14-Day forecast horizons obtained via different ANN constructions using Rolling-Window Validation. First ANN configuration has *null* hidden layers; the second and third has a single hidden layer and the last has two hidden layers.

$0.5 \leq H < 1$ the traffic is self-similar and if $H < 0.5$ the traffic is not self-similar.

In this paper, we use R/S method (also called as Rescaled Adjusted Range Statistics) to calculate H, where

$$E[\frac{R(n)}{S(n)}] = c \times n^H. \tag{1}$$

In the above equation; n is the discrete-time index by which the time-series data is split; $R(n)$ is the amplitude of split time-series as $R(n) = \max\{0, y(1), \ldots, y(n)\} - \min\{0, y(1), \ldots, y(n)\}$; and $S^2(n)$ is the time-series' variance. H is a numerical exponent of the $R(n)/S(n)$ expectation value of that is manifested dependent on a power-law.

According to our calculations, first PoP bandwidth data *is* self-similar with $H = 0.8982$ and that of second is also self-similar with $H = 0.9170$. Following our argument we first apply SARIMA method to predict bandwidth usage.

C. Seasonal Auto-Regressive Integrated Moving Averages

SARIMA method incorporates Auto-Regression (AR) of the time-series and cumulative Moving Average (MA) of the modelling errors. The Integrated (I) part differences the time-series into stationary and a seasonal (S) component is also included. Hyperparameters can be written compactly as SARIMA$(p,d,q)(P,D,Q)_s$ where p/P is trend/seasonal AR-order, d/D is trend/seasonal differencing order, q/Q is trend/seasonal MA-order and s is the seasonal index. In terms of univariate structure model, SARIMA can be formulated as

$$\phi_p(L)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{2}$$

where ϕ, and Φ are weights of the trend and seasonal AR-terms; θ, and Θ are weights of trend and seasonal MA-terms; ∇^d is trend-differencing operator, ∇_s^D seasonal-differencing operator.

Trend and seasonal terms are usually determined using ACF and Partial Auto-Correlation Function (PACF) plots in

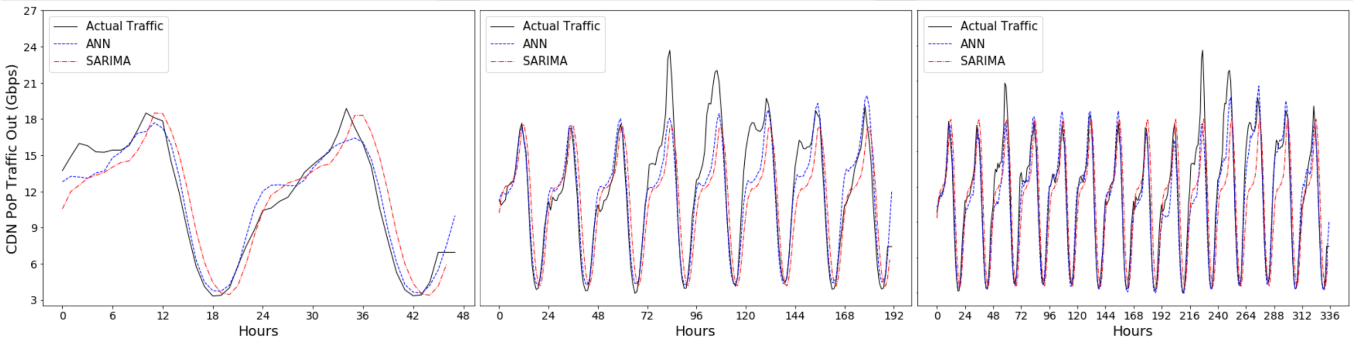


Fig. 4: The first PoP bandwidth forecast for three different forecast horizons with best SARIMA and ANN results. 2-Day forecast is plotted using SARIMA$(2,0,2)(2,0,1)_{24}$ and ANN$(n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{i}/2, n_\mathrm{o})$; 8-Day forecast is plotted with SARIMA$(2,1,2)(2,1,1)_{24}$ and ANN$(n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{o})$; also, 14-Day forecast is plotted SARIMA$(1,0,2)(2,1,2)_{24}$ and ANN$(n_\mathrm{i}, n_\mathrm{i}/2, n_\mathrm{i}/2, n_\mathrm{o})$.

	Validation			Test		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
2-Day Forecast						
SARIMA$(2,1,1)(1,0,2)_{24}$	38.759	3.743	0.951	104.84	6.338	0.957
SARIMA$(1,1,2)(2,0,2)_{24}$	36.202	4.365	0.943	102.72	6.348	0.958
SARIMA$(2,1,1)(1,0,1)_{24}$	42.307	4.107	0.946	100.52	6.360	0.960
SARIMA$(2,1,2)(2,0,2)_{24}$	42.146	3.878	0.951	128.62	7.048	0.935
8-Day Forecast						
SARIMA$(2,0,1)(2,1,1)_{24}$	103.18	7.889	0.763	177.52	10.448	0.902
SARIMA$(2,0,1)(2,0,1)_{24}$	95.531	6.641	0.827	199.50	11.293	0.872
SARIMA$(2,0,1)(2,0,2)_{24}$	100.04	7.379	0.799	213.70	11.727	0.859
SARIMA$(1,0,1)(2,0,2)_{24}$	108.25	8.221	0.772	223.54	12.325	0.846
14-Day Forecast						
SARIMA$(2,0,1)(1,1,2)_{24}$	42.336	3.612	0.982	190.01	10.534	0.895
SARIMA$(2,0,2)(1,1,2)_{24}$	44.915	3.411	0.983	192.01	10.567	0.893
SARIMA$(2,1,2)(1,1,1)_{24}$	45.801	3.918	0.980	191.08	10.673	0.894
SARIMA$(2,0,2)(1,1,2)_{24}$	44.184	3.324	0.983	204.51	11.279	0.878

TABLE III: For the second PoP, most accurate forecasts for 2-Day, 8-Day, and 14-Day forecast horizons obtained via SARIMA hyperparameter grid-search using Walk-Forward Validation. Inspecting Auto-Correlation and Partial Auto-Correlation Function plots, the hyperparameter grid search is done with $p/P, d/D, q/Q = \{1,2\},\{0,1\},\{1,2\}$ for $s = 24$.

	Validation			Test		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
2-Day Forecast						
ANN$(\mathbf{n_i}, \mathbf{n_o})$	209.85	12.929	0.854	316.81	19.821	0.655
ANN$(n_i, 16, n_o)$	210.12	13.431	0.853	333.31	22.241	0.618
ANN$(n_i, n_i/2, n_o)$	213.67	13.194	0.848	348.35	22.895	0.583
ANN$(n_i, n_i/2, n_i/2, n_o)$	211.05	13.157	0.852	341.77	22.459	0.599
8-Day Forecast						
ANN$(\mathbf{n_i}, \mathbf{n_o})$	320.38	20.495	0.663	227.72	12.552	0.828
ANN$(n_i, 16, n_o)$	321.39	20.129	0.661	266.98	15.274	0.764
ANN$(n_i, n_i/2, n_o)$	322.43	20.284	0.659	229.96	13.122	0.825
ANN$(n_i, n_i/2, n_i/2, n_o)$	340.67	20.178	0.619	252.28	15.408	0.789
14-Day Forecast						
ANN(n_i, n_o)	257.29	16.022	0.803	178.22	16.059	0.906
ANN$(n_i, 16, n_o)$	285.81	16.944	0.758	227.95	13.478	0.846
ANN$(\mathbf{n_i}, \mathbf{n_i/2}, \mathbf{n_o})$	253.46	15.007	0.809	183.43	9.951	0.899
ANN$(n_i, n_i/2, n_i/2, n_o)$	280.95	16.369	0.766	193.69	11.009	0.888

TABLE IV: For the second PoP, most accurate forecasts for 2-Day, 8-Day, and 14-Day forecast horizons obtained via different ANN constructions using Rolling-Window Validation. First ANN configuration has *null* hidden layers; the second and third has a single hidden layer and the last has two hidden layers.

which time-series are visualized in terms of ascending order of lags. In Section V, we will get back to methods that we use o determine SARIMA hyperparameters.

In order to determine SARIMA hyperparameters p, d, q, P, D, Q, s; we first graph ACF and PACF plots of the validation sets. For both datasets, the significant lags beyond 95% confidence interval in the ACF lays for the first and second lags so we conclude that $(q = 1, 2)$. Also, significant PACF lags lay in the first and second lags so similarly, we conclude $p = 1, 2$. We then move on the determine whether validation sets are stationary; if the series are stationary then $d = 0$ implying differencing need not be made. We utilize a Dickey-Fuller Test to calculate MacKinnon's p-value to determine whether the validation set has a unit root in the AR series. If $p < 0.5$, any time-series are *stationary* hence differencing is not required. For the validation set Dickey-Fuller Test gives $p = 0.4701$; a result that is too close to the 0.5 threshold; therefore we apply a single order of differencing next to *null* differencing; *i.e.* $d = 0, 1$. After inspecting the ACF plot, we also conclude that our datasets have seasonality as apparent from its oscillating behavior in Fig. 2 and see that mod$(24)^{\text{th}}$ lags exhibit highest significance in descending order. Therefore, we conclude the seasonal index is $s = 24$. Since seasonal parameters should be in the range of trend parameters, we construct a Hyperparameter Grid-Search using Walk-Forward Validation Method for $p/P, d/D, q/Q = \{1,2\},\{0,1\},\{1,2\}$. From 64 possible configurations; best hyperparameter configurations for the PoP datasets Validation and Test sets are shown in Table I and III for the first and second PoPs with 2-Day, 8-Day and 14-Day Forecasts. Before elaborating on our results we continue to present our ANN results either.

D. Artificial Neural Networks

ANNs are a subset of Machine-Learning Algorithms that are non-linear constructs. ANNs must contain at least two

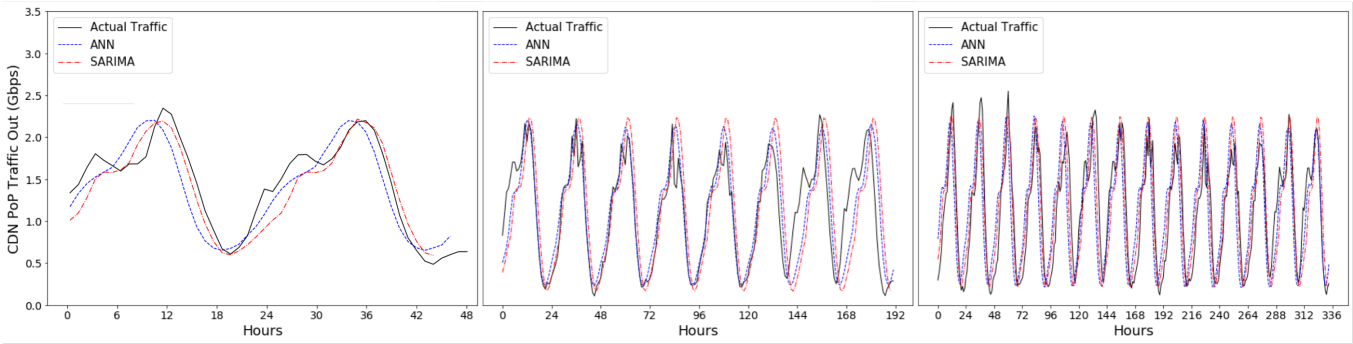


Fig. 5: The second PoP bandwidth forecast for three different forecast horizons with best SARIMA and ANN results. 2-Day forecast is plotted using SARIMA$(1,1,2)(2,0,2)_{24}$ and ANN(n_i, n_o); 8-Day forecast is plotted with SARIMA$(2,0,1)(2,1,1)_{24}$ and ANN(n_i, n_o); also, 14-Day forecast is plotted SARIMA$(2,0,1)(1,1,2)_{24}$ and ANN$(n_i, n_i/2, n_o)$.

layers: The input-layer and output-layer. Adding one-to-many hidden layers in between the input- and output-layers improves the tunability of the underlying non-linear ANN function to present more accurate results. We construct four different configurations of ANNs to forecast CDN PoP bandwidth usage with Validation and Test sets. We use Rolling-Window Validation to train our ANNs and use cumulative errors of performance metrics for back-propagation. In the first configuration, we have as many input neurons (n_{o}) as the number of data points in the Validation/Test set and output neurons (n_{o}) change depending on the 2-Day, 8-Day, and 14-Day forecast horizon data points –where there are no hidden layers in the ANN. The remaining three different ANNs also employ as many input/output neurons as the Validation/Test set but with hidden layers. Second ANN configuration has a single hidden layer with 16 neurons. Third ANN configuration has a hidden layer that is half the number of the input layer, *i.e.* $n_{\mathrm{i}}/2$. Last and fourth ANN configuration has two hidden layers with $n_{\mathrm{i}}/2$-many neurons on each layer. Most accurate ANN configurations for the Validation and Test sets are shown in Table II and IV for the first and second PoPs with 2-Day, 8-Day and 14-Day Forecasts.

IV. Discussion

We now proceed to elaborate on our numerical results in Tables I-IV. Our results can be generalized to CDN PoPs as they belong to the same infrastructure. First; we see that SARIMA hyperparameter search is a good method to obtain more accurate results, and PoP bandwidths can be more accurately predicted using this method. Using a single set of SARIMA hyperparameters would give considerably accurate results; however, the hyperparameter search leads to a better function-fit. On the other hand; ANNs with more hidden layers does not necessarily presents more accurate results. Therefore we found out that both a SARIMA hyperparameter search and adding zero-to-two hidden ANN layers should be utilized for more reliable PoP bandwidth forecasts.

Now we continue to analyze the effect of the forecast horizon. Most reliable predictions for three different forecast horizons are plotted in Fig. 4 for the first PoP and Fig. 5 for the second PoP. Inspecting our results laid out in Tables I-IV; we show that SARIMA presents the most accurate results for 2-Day Forecasts. Four best SARIMA results with different hyperparameter configurations are all superior to all of the ANN configurations in RMSE, MAPE and R^2 error metrics. For the first PoP in Fig. 4; 2-Day Forecast predicted with SARIMA$(2,0,2)(2,0,1)_{24}$ and ANN$(n_{\mathrm{i}}, n_{\mathrm{i}}/2, n_{\mathrm{i}}/2, n_{\mathrm{o}})$ are plotted and SARIMA has the most accurate prediction with RMSE 1035.2, MAPE 7.213% and 0.955 R^2 score. Similarly; we plotted 2-Day forecasts of the second PoP using SARIMA$(1,1,2)(2,0,2)_{24}$ and ANN$(n_{\mathrm{i}}, n_{\mathrm{o}})$ in Fig 5., and SARIMA, again presents a more accurate prediction with RMSE 104.84, MAPE 6.348% and 0.958 R^2 score.

Expanding the forecast horizon to 8 days; we see that SARIMA and ANN results give comparable forecasts. For the first PoP, the most accurate 8-Day forecast is a single hidden layer ANN with RMSE 1888.9, MAPE 11.324% and 0.863 R^2 score. On the other hand, the most accurate for the second PoP is that of SARIMA with RMSE 177.52, MAPE 10.448% and 0.902 R^2 score. Therefore; for 8-Day forecasts, it is better to compute both forecasts as we found neither of the forecast methods to be superior to one another.

When the forecast horizon is expanded to 14 days; we found most accurate results using ANNs. For the first PoP, two hidden layer ANN gives forecast with RMSE 1384.1, MAPE 8.308% and 0.918 R^2 score. for the second PoP single hidden layer ANN gives the most accurate forecast RMSE 183.43, MAPE 9.951% and 0.899 R^2 score.

To sum up; our analyses include three different forecast horizons and we find that for 2-Day forecasts SARIMA is superior and for 14-Day forecasts ANN produces more accurate results. For the 8-Day forecasts, our results show that neither of the methods are better.

V. Conclusion

In this paper, we compared SARIMA and ANN for CDN traffic prediction. Using our results, CDNs can predict their PoP bandwidth usages to better their resource allocation, to shorten their end-user content delivery latencies and can predict their server allotments.

Bandwidth usage forecast for Content Delivery Networks has never been studied in the literature within a networking perspective. In this paper, our approach accounts for the bandwidth usage directly between the PoP and the end-user, so fluctuations in internet traffic due to various reasons are also tacitly incorporated in our results.

Our forecasting results takes into account the statistical characteristics of the CDN PoP traffic. Also, SARIMA prediction methods has never been applied from a CDN perspective; whereas the strongest suit for our new approach is to diminish the traffic fluctuations caused by local ISPs as CDN PoP bandwidth is measured between the CDN and the end-user.

We consider the results of this paper to be easily implementable to current CDN architectures. CDNs can use our results to redirect their current traffic at different PoPs, depending on the different forecast horizons.

References

[1] P. Sekwatlakwatla, M. Mphahlele, and T. Zuva, "Traffic flow prediction in cloud computing," in *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, Nov 2016, pp. 123–128.

[2] B. L. Dalmazo, J. P. Vilela, and M. Curado, "Predicting traffic in the cloud: A statistical approach," in *2013 International Conference on Cloud and Green Computing*. IEEE, Sep. 2013, pp. 121–126.

[3] B. L. Dalmazo, J. P. Vilela, and M. Curado, "Online traffic prediction in the cloud: a dynamic window approach," in *2014 International Conference on Future Internet of Things and Cloud*. IEEE, 2014, pp. 9–14.

[4] R. Madan and P. S. Mangipudi, "Predicting computer network traffic: A time series forecasting approach using dwt, arima and rnn," in *2018 Eleventh International Conference on Contemporary Computing (IC3)*. IEEE, Aug 2018, pp. 1–5.

[5] L. Tang, S. Du, and S. Ji, "Forecasting network traffic at large time scale by using dual-related method," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE, Aug 2016, pp. 1336–1340.

[6] Y. Wang, Y. Liu, and Y. Gan, "Research on combination network traffic forecasting model," in *2018 IEEE International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, Nov 2018, pp. 311–314.

[7] M. Li and S. Lim, "Modeling network traffic using generalized cauchy process," *Physica A: Statistical Mechanics and its Applications, Elsevier*, vol. 387, no. 11, pp. 2584 – 2594, 2008.

[8] M. Papadopouli, E. Raftopoulos, and H. Shen, "Evaluation of short-term traffic forecasting algorithms in wireless networks," in *2006 2nd Conference on Next Generation Internet Design and Engineering, 2006. NGI '06.* IEEE, April 2006, pp. 8 pp.–109.

[9] P. Cortez, M. Rio, M. Rocha, and P. Sousa, "Internet traffic forecasting using neural networks," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, July 2006, pp. 2635–2642.

[10] G. Zhang, B. Patuwo, and M. Y. Hu, "A simulation study of artificial neural networks for nonlinear time-series forecasting," *Computers Operations Research, Elsevier*, vol. 28, no. 4, pp. 381 – 396, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305054899001239

[11] R. Boutaba, M. A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. M. Caicedo, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," *Journal of Internet Services and Applications, Springer*, vol. 9, no. 1, p. 16, 2018.

[12] J. Ruiz Aguilar, I. Turias, and M. Jimenez Come, "Hybrid approaches based on sarima and artificial neural networks for inspection time series forecasting," *Transportation Research Part E: Logistics and Transportation Review, Elsevier*, vol. 67, p. 1–13, 07 2014.

[13] W. Pannakkong, V.-N. Huynh, and S. Sriboonchitta, "A novel hybrid autoregressive integrated moving average and artificial neural network model for cassava export forecasting," *International Journal of Computational Intelligence Systems, Atlantis Press*, vol. 12, p. 1047, 09 2019.

[14] D. Alencar, C. Affonso, R. Oliveira, and J. C. Reston Filho, "Hybrid approach combining sarima and neural networks for multi-step ahead wind speed forecasting in brazil," *IEEE Access*, vol. PP, pp. 1–1, 10 2018.

Optimization of Thick BoR Monopole Antennas Using Differential Evolution

Marko Radović
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
marko.radovic@student.um.si

Gorazd Lešnjak
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
gorazd.lesnjak@um.si

Peter Kitak
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
peter.kitak@um.si

Peter Planinšič
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
peter.planinsic@um.si

***Abstract*—Two electrically thick body BoR monopole antennas are proposed with enhanced antennas gain performances. Both antennas are designed above infinite ground plane, fed with coaxial cable and to operate in frequency range from 0.3 GHz to 1.2 GHz. Voltage standing wave ratio and realized gain characteristics are observed and studied. The enhancement was obtained using optimization with Differential Evolution (DE) algorithm. Proposed antennas have better gain performances in comparison to reference conical cylindrical antenna. Results obtained for second antenna show improved impedance bandwidth performances in the upper part of frequency band in comparison to reference antenna.**

Keywords— BoR monopole antenna, antennas gain, optimization, Differential Evolution

I. INTRODUCTION

The interest in developing and improving rotationally symmetrical compact Wide Band antennas is known for more than half of century [1], [2]. Nowadays wireless Wide Band and Ultra-Wide Band (UWB) communications with higher and higher carrier frequencies, oft use compact and small printed board circuits (PCB) antennas or even antennas integrated on RF chip. However, compact, small size and physical robust rotationally symmetrical antennas are still very attractive, due to their omni-directional radiation , good antennas gain and other properties. They are appropriate for difficult environments such as for example appear in modern mobile wireless vehicle and military communication applications.

Examples of compact and robust symmetrical antennas are thick and low profile Body of Revolution (BoR) antennas.

Analysis and synthesis of electrically thick BoR antennas were given by Djordjević, Dragović and Popović in [2], where a method for computer-aided analysis and synthesis of electrically thick BoR antennas was presented with experimental results for verifying their method. Among their antennas examples there was electrically thick BoR monopole antenna referred to as cylindrical-conical monopole antenna which we use as a reference antenna (Reference model) in this paper.

Another approach in designing BoR monopole antennas was the method presented for Ultra-Wide Band (UWB) their performance optimization using random walk profile, introduced by Zhao [3]. Low-profile UWB inverted-hat (IHA) monopole antenna proposed by Zhao [4] was designed to operate with frequencies from 50 MHz to 2 GHz, especially appropriate for small unmanned aerial vehicles. In [5] Zhao introduced Genetic Algorithms (GA) for optimizing Low-profile UWB inverted-hat (IHA) monopole antenna. Miniature compact axisymmetric resonant lens antenna with improved directivity in *Ka*-Band was presented in [6]. A comprehensive overview of Wideband and UWB Antennas for Wireless Applications was presented in [7]. Analysis methods and applications of low-profile natural and metamaterial antennas was given in [8].

Modern stochastic bioinspired optimization algorithms enable efficient optimization of complex nonlinear problems, and were oft used for optimizations of different types of antennas. An overview of evolutionary algorithms applied for antennas and propagation was made in [9].

One of the popular bioinspired algorithm in the family of Evolutionary Algorithms (EA) is Differential Evolutionary (DE) algorithm. We used classic DE algorithm [10] in our study of BoR monopole antennas design optimization in combination with modeling and analyzing antennas with 3D Electromagnetic (EM) Solver tool WIPL-D Pro [11]. WIPL-D Pro is very efficient program for fast and accurate analysis of metallic and/or dielectric/magnetic structures. Classic DE algorithm is known as very simple and efficient. Also BoR antennas are relative simple for analysis and design. However, the motivation in our study was to investigate if some further improvements in designing of compact BoR antennas using (classic) DE optimization algorithm can be achieved.

We proposed two types of new thick BoR-monopole antennas, BoR-model$_1$ and BoR-model$_2$, respectively. They were designed by comparing of their electrical characteristics (gains) with characteristic of Reference model (antenna). Geometry of triple-ellipse IHA-antenna presented by Zhao was used for constructing the geometry of BoR-model$_1$ [3]. The second BoR-model$_2$ is based on geometry of optimized UWB BoR monopole antenna #2, also presented by Zhao [4]. All antennas have been modelled in WIPL-D Pro v11 assuming that they are made of perfect electrical conductor and matched to 50 Ω coaxial line.

The analysis was carried out in the frequency range from 0.3 GHz to 1.2 GHz. Simulated gain performances of both new monopole antennas show clearly improvement in the frequency range from 0.45 GHz to 1.1 GHz in comparison to

Reference-model. Results for voltage standing wave ratio (VSWR) show also impedance bandwidth (matching) improvement for BoR- model$_2$ in upper part of the frequency band from 0.9 GHz to 1.15 GHz.

II. Methodology and Results

A. Initial Geometries of Anntenas Models

Fig. 1 shows that all three antennas (BoR-model$_1$, BoR-model$_2$, and the Reference model) consists of three geometrical parts with lengths l_1, l_2 and l_3, respectively. The radius of the middle cylindrical part is denoted as R. Reference antenna consists of two cones and a cylinder.

Geometry shape modifications of proposed antennas in comparison to Reference antenna was done only at the part that is connected to the coaxial line. All three antennas were modeled as placed perpendicular above infinite ground plane and fed by a coaxial line with characteristic impedance 50 Ω. The radius a in Fig. 1 is the radius of inner coaxial line and was set to 3.3 mm, while outer radius of coaxial line b was set to 7.59 mm, the same for all three antennas.

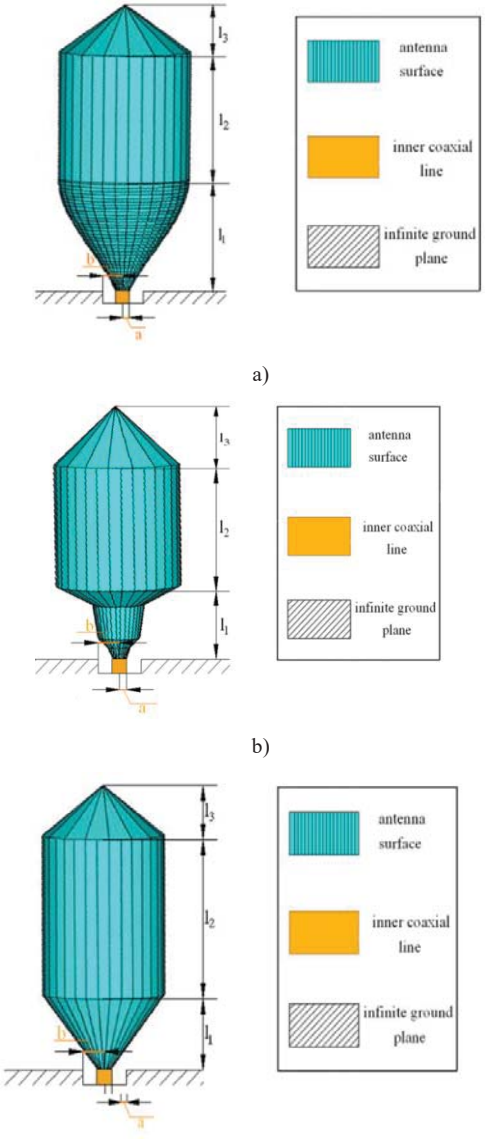


Fig. 1. Geometrical models of electrically thick BoR monopole antennas: a) BoR-mode$_1$, b) BoR-model$_2$ and c) Reference model

B. Short Describtion of DE Algorithm

Differential evolution algorithm was first proposed by Storn and Price in [10] and has later received a great attentions and became popular and powerful optimization tool in engineering in the family of evolutionary algorithms [12]. Up today there were many enhancement of classical DE algorithm [13].

Basic principle of DE is as follows. It is stochastic, population based algorithm for global optimization of nonlinear, non-differentiable, noisy, and flat functions with multiple variables (parameters), with constraints and many local minima. It starts with a uniformly random set of candidate solutions from the feasible search volume. In every iteration, e. g. generation of algorithm it has the same computational steps as a standard Evolutionary Algorithm (EA). However, DE differs significantly from EA algorithm in that, that it mutates the base vectors, e. g. secondary parents with scaled differences of the important members from the current population. This property is called self-referential mutation. While EA and some Genetic Algorithms (GA) require the adaptation for each variable over iterations, the canonical DE requires only the adaptation of a single relative scale factor for all variables. Unlike several other EA techniques, the basic DE is a very simple algorithm whose implementation requires only a few lines of code in any standard programming language, like Matlab, used in our case (Fig. 2). In addition, the basic DE requires only three control parameters, the scale factor, the crossover rate and the population size, respectively (Fig. 3). But surprising, DE has a very good performance while optimizing a wide variety of multi variable objective functions in terms of final accuracy, computational speed, and robustness.

As mentioned, when constructing first geometrical part of two new antennas we used ideas presented by Zhao for constructing antenna profiles of IHA and UWB BoR monopole antenna, respectively. Profiles of first geometrical part of proposed two new antennas and the reference antenna are presented in Fig. 4.

For constructing first geometrical part of BoR-model$_1$ we used a profile that is similar to the profile of triple-ellipse IHA antenna, with mathematical description

$$y = \frac{R}{2}\left(1 + cos\frac{\pi n t}{l_1}\right) \qquad (1)$$

where variable t belongs to interval [0, R] and factor n is chosen to get $y = a$ at $t = R$. Expression for n is

$$n = \frac{l_1}{\pi R}\left(arccos\frac{2a}{R} - 1\right) \qquad (2)$$

C. Antennas Designing Optimization Using DE Algorithm

Shape optimization process for both new models was carried out to find four optimal geometrical dimensions R, l_1, l_2 and l_3 to achieve enhanced gain performances of both new antennas. Obtained results for optimized parameters of both new models and the reference antenna are presented in Table 1. Optimized parameters were obtained after 100 iterations of DE algorithm.

The antenna gain is defined as the ratio between the reference power density P of an isotropic radiator, and the power density P_{di} in particular consideration direction. It is usually expressed in dB [14]:

$$G_{dBi} = 10 \cdot log_{10}\left(\frac{P_{di}}{P}\right) \quad (3)$$

In our DE optimization problem, we vary four antennas geometrical parameters l_1, l_2, l_3, and R to maximize G_{dBi} in considered frequency range in comparison to Reference model (antenna). In DE algorithm we minimize the cost function which is the inverse of the sums of absolute gain differences of considered new designed antenna model and Reference model in 31 points.

For each obtained new antenna we considered also Voltage Standing Wave Ratio (VSWR), which is a measure of power of the reflected wave traveling back to the transmitter via connecting feeder cable, because of not appropriate adaptation of feeder characteristic impedance to the antenna input impedance. The VSWR is expressed as [14]:

$$VSWR = \left(\frac{Vmax}{Vmin}\right) \quad (4)$$

where V_{max} and V_{min} are maximum and minimum voltage amplitudes of waves on feeder cable. A good much will result in a value of approximately 1.2. In the case of total matching its value is 1. The relation of VSWR to impedances can be written as [15]:

$$VSWR = \frac{1+|\Gamma|}{1-|\Gamma|} \quad (5)$$

where Γ is voltage reflection coefficient at the input terminals of the antenna. It is defined as:

$$\Gamma = \frac{Z_{in}-Z_0}{Z_{in}+Z_0} \quad (6)$$

where Z_{in} is antenna input impedance and Z_0 is feeder cable (transmission line) characteristic impedance . In the case of total matching, Z_{in} is equal to Z_0 (in our case 50 Ω). Then the value of Γ is 0 and VSWR is 1. VSWR is related to complex valued S-parameter S_{11}, which is equal to reflection coefficient Γ and is usually plotted in Smith chart [15].

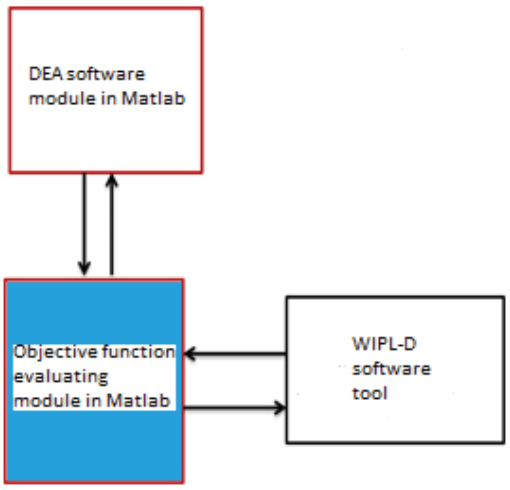


Fig. 2. Sotware tools used in our antennas design approach

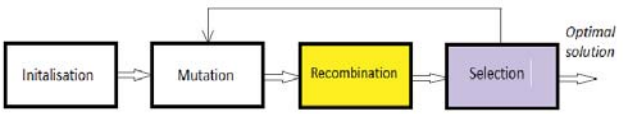


Fig. 3. Principal block scheme of DE algorithm

TABLE I. OPTIMIZED ANTENNA PARRAMETERS WITH DE

Parameter (mm)	l_1 (mm)	l_2 (mm)	l_3(mm)	R (mm)
BoR-model 1	60.62	72.61	23.55	32.41
BoR-model 2	44.12	80.67	34.24	33.00
Reference model	50.00	87.00	3.00	33.00

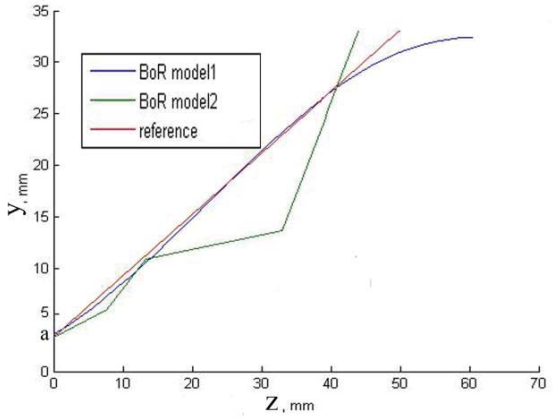


Fig. 4. The antenna 2-D profiles for first geometrical part of reference and proposed two antennas.

2-D profile of the first geometrical part for BoR-model$_2$ was composed from four line segments similar to optimized antenna #2 proposed by Zhao. In optimization process of UWB IHA monopole antenna introduced by Zhao line segments of 2-D antenna profile were functionally dependent to its length and angle. In our case optimization process for 2-*D* profile of BoR-model 2 was performed only along the z–axis as shown in Fig. 4.

Two proposed new electrically thick BoR monopole antennas together with the reference antenna were simulated using 3D EM Solver WIPL-D Pro v11. Results obtained for VSWR in frequency range from 0.3 GHz to 1.2 GHz for all three antennas are presented in Fig. 5, VSWR in chartezian coordinates and complex valued parameter S_{11} in Smith chart. It can be seen that graph for the BoR-model$_1$ almost completely overlaps graph of Reference antenna. Improvement for impedance bandwidth of BoR-model$_2$ is visible in upper part of frequency range, from 0.9 GHz to 1.15 GHz. Maximum achieved improvement for model$_2$ of 12.39 % for impedance at frequency 1.08 GHz is designated with green markers in Fig. 5.

Fig. 6 shows simulated results for gain in the same frequency range for all three antennas. It is observed that both proposed antennas express enhanced gain performances in comparison to the reference antenna. This is evident in

frequency range from 0.45 GHz to 1.1 GHz, while this is less noticeable in the rest of frequency band.

Maximum enhanced gain performances for BoR-model$_1$ in comparison to Reference antenna are designated with blue markers at frequency 0.72 GHz. Maximum enhanced realized gain performances of BoR-model$_2$ are designated with green markers at frequency 0.96 GHz. Maximum enhanced gain 0.23 dBi for BoR-model$_1$ is reached at frequency point 0.72 GHz while maximum enhanced realized gain 0.2 dBi for BoR-model$_2$ is attained at frequency point 0.96 GHz.

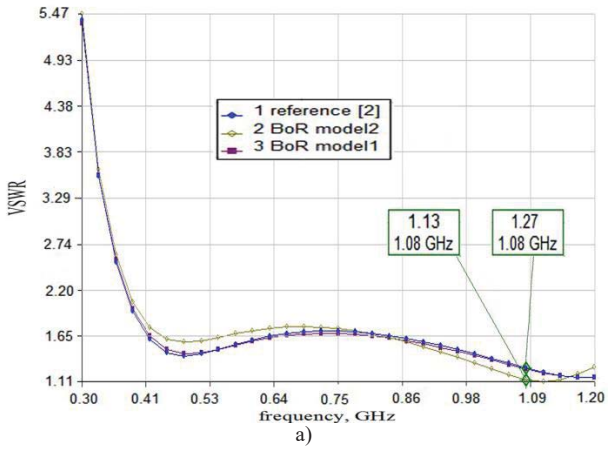


Fig. 5. a) Simulated VSWR variation of three antennas vs frequency obtained with WIPL-D Pro v11., b) Parameter S_{11} ploted in Smith chart.

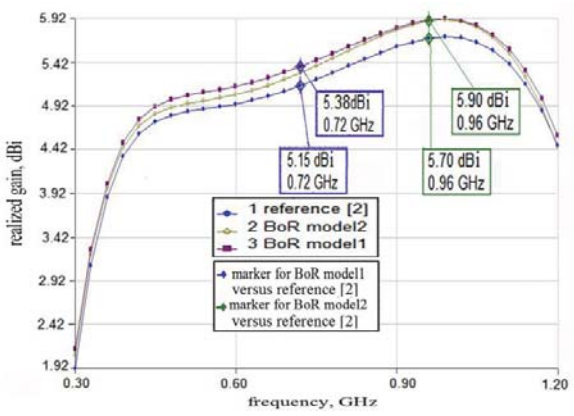


Fig. 6. Simulated gain variation of three antennas, reference and proposed, vs frequency, obtained with WIPL-D Pro v11.

III. Conclusions

Two novel electrically thick monopole antennas, BoR-model$_1$ and BoR-model $_2$ were introduced. Electromagnetic simulations for both new antennas and Reference antenna have been done in frequency range from 0.3 GHz to 1.2 GHz. Results of simulations in proposed frequency band and configurations for all three antennas show enhanced gain performances for both new antennas in comparison to Reference antenna. This is clearly evident in frequency range from 0.45 GHz to 1.1 GHz. Improved impedance bandwidth performances of BoR-model$_2$ antenna has been also achieved in frequency range from 0.9 GHz to 1.15 GHz with respect to conical-cylindrical monopole Reference antenna.

Acknowledgment

Authors like to thank prof. dr. Igor Tičar for supporting this work.

References

[1] M. G. Andreasen, "Scattering from bodies of revolution", *IEEE Trans. Antennas Propag.*, 1965, 13, (2), pp. 303–310.

[2] A. R. Djordjević, M. B. Dragović, B. D. Popović, "Analysis of Electrically thick antennas of revolution", Proc of 3rd ICAP, Norwich , April 1983, pp. 31–35

[3] J. Zhao, T. Peng,T., C.-C. Chen, J. L. Volakis, "Low-profile ultra-wideband inverted-hat monopole antenna for 50 MHz–2 GHz operation", Electronics Letters, 2009, vol. 45, no 3, pp. 142-144, July 2011, pp. 3295–3296.

[4] J. Zhao, D. Psychoudakis, C.-C. Chen, J. L. Volakis, "Ultra-wideband performance optimization of a body-of-revolution monopole antenna", IEEE (APSURSI) Int. Sym. Antennas Propag., 2008, Spokane, WA

[5] J. Zhao, D. Psychoudakis, C.-C. Chen, and J. L. Volakis, Design Optimization of a Low-Profile UWB Body-of-Revolution Monopole Antenna, *IEEE Transaction on Antennas and Propagation*, vol. 60, no. 12, December, 2012, pp. 5578-5586.

[6] A. Rolland, M. Ettorre, A. V. Boriskin, L. L. Coq, and R. Sauleau, Axisymmetric Resonant Lens Antenna With Improved Directivity in Ka-Band, *IEEE Antennas and Wireless Propagation Letters*, vol. 10, 2011, pp. 37-40.

[7] R. Cicchetti, E. Miozzi, and O. Testa, Wideband and UWB Antennas for Wireless Applications:A Comprehensive Review, International Journal of Antennas and Propagation Volume 2017, Article ID 2390808, 45 pages, Hindawi Publishing Corporation, https://doi.org/10.1155/2017/2390808

[8] H. Nakano, *Low - Profile Natural and Metamaterial Antennas: Analysis Methods and Applications*, Willey, 2016.

[9] S. K. Goudos, C. Kalialakis, and R. Mittra, Evolutionary Algorithms Applied to Antennas and Propagation: A Review of State of the Art, *International Journal of Antennas and* Propagation, Volume 2016, Article ID 1010459, 12 pages, Hindawi Publishing Corporation, http://dx.doi.org/10.1155/2016/1010459.

[10] R. Storn, K. Price, "Differential Evolution: A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces", J. Global Optimization, 1997, vol. 11, no 44, pp. 341-359.

[11] Wipl-d Pro v11 simulator "Wipl-dpro 3D electromagnetic solver for fast and accurate analysis of arbitrary metallic and dielectric/magneticstructures", http://www.wipl-d.com/products.php?cont=wipl-d-pro, accessed July 2015.

[12] G. C. Onwubolu, B. V. Babu, *New Optimization techniques in Engineering*, Springer, 2004.

[13] D. Swagatam S. S. Mullick, and P. N. Suganthan," Recent Advances in Differential Evolution – An Updated Survey", *Swarm and Evolutionary Computation*, February 2016.

[14] T. Manning, *Microwave Radio Transmission Design Guide*, Artech House, 2006.

[15] C.A. Balanis, *Antenna Theory. Analysis and Design.*Wiley-Interscience, 2005.

DVFS Technique on a Zynq SoC-based System for Low Power Consumption

Marsida Ibro
Faculty of Information Technology
Aleksandër Moisiu University
Durres, Albania
marsidaibro@uamd.edu.al

Galia Marinova
Faculty of Telecommunications
Technical University of Sofia
Sofia, Bulgaria
gim@tu-sofia.bg

Abstract - This paper analyses the impact on power consumption when the Dynamic Voltage and Frequency Scaling (DVFS) technique is implemented on a SoC Zynq 7000 device. The usage of the DVFS technique allows the hardware IP Core design to reduce the typical power consumption. The main concern is about static and dynamic power consumption reduction by selecting the right CPU clock frequency using the DVFS technique. Several wide-ranging power consumption reduction techniques usually disregard the operating characteristics. Subsequently, we present in this paper, not only the hardware design and the operating characteristics but also the needed measurements for different operation modes to enhance the design for power consumption efficiency. Most of the experiments are conducted on the processing unit, whereas the CPU clock frequency and input voltage for Programmable Logic (PL) systems are altered. The empirical results from the application of the DVFS technique indicate that the worst scenario is when the input voltage supply for PL and CPU clock frequency have the maximum values. The best scenario for this design is when the CPU clock frequency is highest and the input voltage supply for PL is minimal, where the measurements for power consumption, especially for dynamic power consumption show that the value is reduced by additional 3%.

Keywords - Zynq 7000 AP - SoC, DVFS, low power consumption

I. INTRODUCTION

Field Programmable Gate Arrays (FPGAs) devices have always been considered as a good choice for the digital design of electronic hardware systems. FPGAs are largely counted as digital devices which consume higher power compared to Application-Specific Integrated Circuits (ASICs). But FPGAs are more flexible due to reconfiguration property and efficient hardware reuse. Power consumption in FPGAs devices has been expected to be not as good as than in Application-Specific Integrated Circuits (ASICs) and therefore they are limited in applications with energy constrain [1]. Currently, many FPGA manufacturers on the market are offering devices with different technologies and the recent 28 nm FPGAs devices consume about 50% lower power than the previous generations. The advance of FPGAs device fabrication is making possible the application of different techniques depending on the hardware architecture of the system implemented.

Dynamic Voltage and Frequency Scaling (DVFS) technique is effective in power consumption reduction of both dynamic and static power on FPGAs devices by just applying the right voltage and frequency value. Utilizing the DVFS technique for such devices support voltage and frequency adaptation depending on the volume of processing, fabrication technology and operating conditions [2]. For example, complex digital processing systems are the right resolution for the increasing challenges between performance and power consumption for nowadays applications. Big data analysis for real-time computing with low-delay, requires an instant response and the adapt scenario configuration with maximum voltage and frequency, is the most suitable. Clock gating usually can be used to control and lower the device temperature when recent operations are not implicated and the transition to the active state is possible within a single clock cycle [3].

In most of the literature reviews, we find that many kinds of research are done in the optimization field where the usage of power consumption reduction for FPGA devices mostly takes into consideration the DVFS technique. In many approaches for power consumption reduction, researches have been focused on developing recent FPGA device architectures that employ multi-threshold voltage techniques and energy at gate-level techniques [4–8].

Other approaches have been persuaded for mainly adjusting the mapping process and place and route algorithms will provide a considerable power consumption reduction [9–11]. This research gives an overview and emerges the necessity of FPGA manufacturers and layout tools to adopt innovative platforms and design environments. In these papers, the authors present the dynamic voltage scaling (DVS) technique as an effective solution used for power consumption management in SoC FPGA devices.

In the DVS technique, the FPGA device input voltage is controlled by the management of the power supply and by changing the internal voltage of the device. This work presents mainly the concept of DVS power-saving competence in commercially available FPGAs and consequently does not focus on implementation strategies to deliver power optimization and cost-effectiveness. A comparable approach for this technique, based on delay has been shown in the work done by Nabina and Nunez-Yanez [12].

The DVFS technique is mostly proposed to minimize the power consumption of an FPGA-based processing unit, primary by adapting the right value of the input voltage, then defining the appropriate frequency at which it will operate. Once again, in this approach, initially it is necessary to define the critical elements and then the delay of the logic circuit is foremost used to trace all the critical points during operation when the voltage and frequency are scaled. Substantially power consumption is measured when the internal supply voltage of PL, V_{CCINT} voltage is decreasing from its value of 1.0 V (maximum permitted value is 1.1 V) to its lower limit of 0.6 V.

The work presented in this article is structured as follows: in section II mostly are explained the key concepts in the architecture of 7-series devices and focus especially in Zynq 7000 SoC; section III provides the methodology used to implement the DVFS technique in the processing system designed. Following, Section IV presents the results from the measurements done while the value of the clock frequency and the value internal supply voltage of Programmable Logic (PL) are varied. The technique implemented is evaluated through computed results in the tool Vivado.

II. HARDWARE IP CORES

FPGA devices are turning out to be a good choice for digital design due to the capability of reconfiguration which can be used for complex systems. Recently, programmable logic devices have been developed in lots of different architectures designed to fulfil further requirements regarding performance. An Intellectual Property (IP) represents a reusable unit of logic developed to provide licensing to multiple vendors. Today IC circuits comprise all system functionalities into a single chip (System on Chip). There are two major types of IP cores: Soft IP that offers RTL synthesis models to create designs at the gate-level netlist and Hard IP, on the other hand, that is based on layout designs in GDS format which cannot be customized.

A. 7-series devices with 28 nm technology

The FPGA devices known as 7-series, which are typically SoCs, are fabricated with 28nm technology to ensure high-performance and low-power consumption. These devices are designed based on architectural and block-level innovations such as a dynamic function of static power reduction like Multi-mode I/O control; Intelligent clock gating; dynamic voltage and frequency scaling [13].

Intellectual property (IP) cores are an appropriate issue for influencing the selection of the FPGA vendors and especially are used for specific complex designs. IP cores are built to ensure all needed tools for integrating complex logic functions in your designs, from high-speed Gigahertz transceivers to digital signal processors known as soft-core processors and embedded ARM system on a chip.

B. Zynq 7000 All Programmable SoC

Zynq™ 7000 device family is offering the new feature of Extensible Processing Platform (EPP). Xilinx presents through hard IP core, Zynq 7000 AP-SoC device the merging together of the characteristics of a high-performance ARM dual-core microprocessor with the 28nm technology. This processor-based architecture ensures the needed flexibility and scalability of an FPGA combined with ASIC-based performance and power consumption for application-specific standard product.

Zynq-7000 All Programmable SoCs are designed by Xilinx with the purpose to design innovative and effective systems based on IP cores. These devices combine a set of functions such as a fast processor system based on two 1GHz ARM® Cortex™-A9 MP Core processors with the industry's fastest and most advanced 28nm FPGA fabric, multiple high-speed transceivers, and on-chip analogue-processing block that incorporates two A/D converters. Commonly, the Zynq-7000 device is composed of the processing system (PS) and the programmable logic (PL). The communication between these units can be realized by the communication interfaces and busses such as General-Purpose Input/Output (GPIO), Advanced eXtensible Interface (AXI), EMIO (Extended Multiplexed Input/Outputs), and DMA (Direct Memory Access). The AXI interconnect interface includes ports such as one master port 64-bit on the PL unit, four master ports with high-performance in the PL unit which can be programmed and optimized for high bandwidth communication from the PL unit to external memories, and four general-purpose ports to access peripherals and registers/memories. The allowed and most appropriate bandwidth for PS-PL units and PS-memory interfaces are given in this paper [14]. The works presented in [15, 16] show further conclusions that support the evaluation of the performance, which is particularly useful in practical designs.

III. METHODOLOGY

New competitive FPGA devices are being designed but the complexity of their architecture makes more difficult the task to provide low-power consumption. For FPGA devices, power consumption is composed of static and dynamic power. These devices are constructed with transistors which should not dissipate power because this will lead to lower values for the static power consumption. Static power consumption is proportional to the static current that flows regardless of gate switching and dynamic power consumption is related to the active current that flows during switching activities. Some other crucial components to take into consideration during the digital design with FPGA devices are the package, frequency of operation, toggle rate, heat sink, fan, the board size, ambient temperature, and resource utilization.

Furthermore, the requirements for the performance of complex systems designs indicate the need for taking into account all the necessary components trade-offs during the design that affect power consumption for a reliable system. As a consequence, power consumption turns out to be a complicated parameter that needs to be estimated and measured. Since there are many different power consumption calculator tools which provide a clear overview of the power consumed by each component and simplify the selection of the best devices based on low power technologies and device dimension. This gives a complete overview of how each element contributes to the overall power consumption and consents the designers to decide exactly how to do the best optimization for reducing the total power consumption.

To design the Processing Unit (PU), we have used the ZedBoard device as a development platform because it offers the integration of processing unit based on ARM Cortex-A9 MP Core dual-core and dual precision floating point. This processing unit uses a memory block BRAM with 4K and ensures the necessary input/output peripherals in the AXI_GPIO block. ARM A9 processors are based on virtual memory and can execute 32-bit instructions. This processing unit is designed to be a general-purpose for data processing. After the design of the Processing Unit (PU) and the process of verification and validation is done through the Vivado ISE design tool (version 2014.2) in Fig. 1.

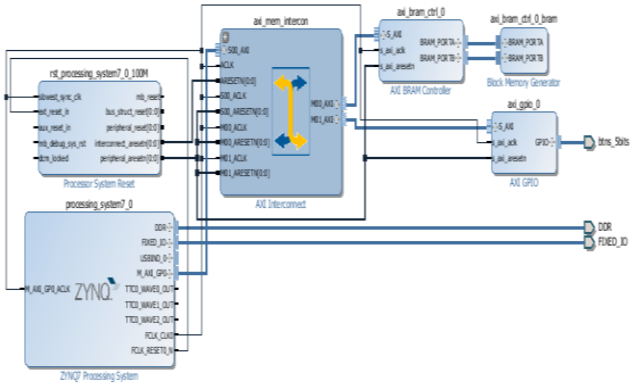


Fig. 1. Block design of Processing Unit (PU)

After the block design has been completed, the Vivado tool is used to compile and it is converted in the VHDL programming language that can be used later to implement the design based on the FPGA logic. This includes several steps like synthesis, simulation, implementation and finally generating the bitstream. Implementation is initiated once the simulation is successful and this process includes logical as well as physical transformations of the design as shown in Fig. 2. Vivado displays the Zynq 7000 device resources allocated (floorplanning), to implement the designed processing unit such as SLICEs, RAMs, clock and DSPs. The figure shows that the resource allocation is optimized enough to make the place and route process more convenient.

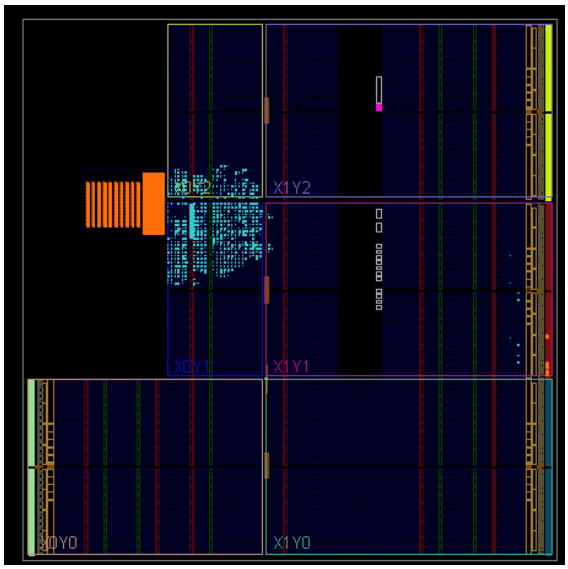


Fig. 2. Floorplannig for the designed Processing Unit (PU)

The DVFS technique is applied to the designed Processing Unit and will support further monitoring of the power consumption of the xc7z020clg484-1 element [17]. This technique augments the optimization for the power consumption of the Zynq 7000 element in the ZedBboard. Dynamic frequency scaling was done by using a reconfigurable clock on the PL unit. The clock frequency is set by the configuration of the registers in the global clock line as shown in the block scheme of Fig. 3.

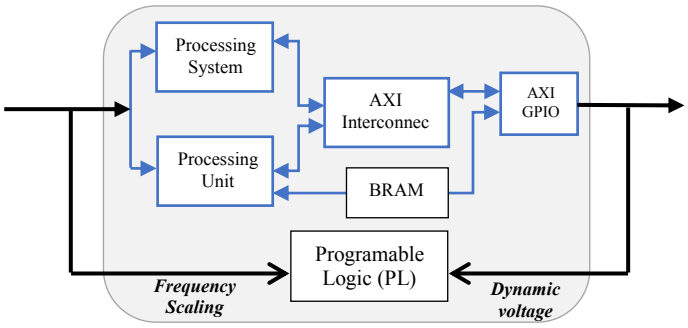


Fig. 3. Block scheme of DVFS for a Zynq-based system

In this work, we have modified and changed the values to measure the power consumption of the Programmable Logic (PL) unit for the purpose to not affect the data processing rate. Dynamic voltage scaling of PL unit, the internal supply voltage V_{CCINT} should be changed carefully, since it may cause permanent damage to the element on the board if the selected voltage exceeds the limits described in the manual [18]. Power management of the PS unit is not advised since the monitoring application will consume some power and will not provide accurate results. The clock input for PL generated by the global clock on the SoC can run at a reduced clock rate by using the internal Phase-Locked Loop (PLL).

IV. Experiment Results

The DVFS technique is used for the ZedBoard because it reduces the power dissipation when the voltage and frequency are changed and are not fully interdependent from each other. If the clock frequency is decreased without changing the internal supply voltage V_{CCINT}, consequently it implies a decrease in power dissipation but may lead to significant changes in static power consumption. If the internal supply voltage V_{CCINT} is decreased without changing the operating frequency may imply power consumption reductions. Therefore, by decreasing internal supply voltage V_{CCINT}, we can reduce power consumption by V^2. The experiments were carried out using Vivaldo and results are analysed from Vivado Report Power and Report Utilization tools.

A. *Scenario 1*

To apply the DVFS technique for the processing unit, we have incremented the CPU clock frequency between the values of 50 to 650 MHz and the internal supply voltage V_{CCINT} of PL unit is fixed at the maximum value of 0.950V. The toggle rate and default switching activity are fixed at a value of 12.5%. The resource utilization is 2.75% for slice LUTs, 1,58 % for slice registers, 1.43% of memory, 2.48 % for I/O and 3.12 % is for clocks. The results obtained are presented in Table I and Fig. 4.

B. *Scenario 2*

To apply the DVFS technique for the processing unit, we have fixed the CPU clock frequency to the maximum value of 650 MHz and the internal supply voltage V_{CCINT} of PL unit is incremented between the values from 0.950V to 1.050V. The toggle rate and default switching activity are fixed at a value of 12.5%. The resource utilization is 2.75% for slice LUTs, 1,58 % for slice registers, 1.43% of memory, 2.48 % for I/O and 3.12 % is for clocks. The results obtained are presented in Table II and Fig. 5.

TABLE I. SCENARIO 1 MEASUREMENTS

Scenario 1	The CPU clock frequency is variable and VCCINT is fixed	
	Static power [W]	Dynamic power [W]
50 MHz	0.132	1.29
150 MHz	0.132	1.292
250 MHz	0.132	1.293
350 MHz	0.133	1.296
450 MHz	0.133	1.297
550 MHz	0.134	1.298
650 MHz	0.136	1.296

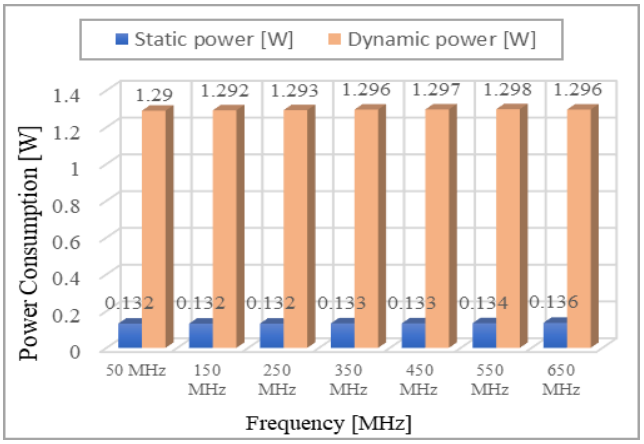


Fig. 4. Frequency scaling and the PL power consumption monitoring

TABLE II. SCENARIO 2 MEASUREMENTS

Scenario 2	The CPU frequency is fixed and VCCINT is variable	
	Static power [W]	Dynamic power [W]
0.950V	0.149	1.004
0.960V	0.149	1.225
0.970V	0.15	1.331
0.980V	0.15	1.335
0.990V	0.151	1.338
1.000 V	0.156	1.341
1.050 V	0.157	1.342

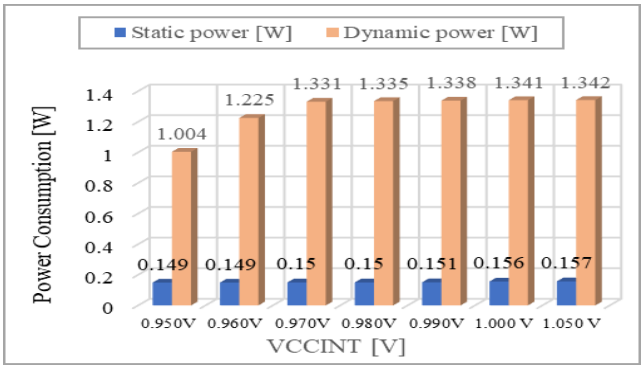


Fig. 5. Dynamic Voltage and the PL power consumption monitoring

V. CONCLUSIONS AND FUTURE WORK

This paper aims to analyze the power consumption in the relationship between the CPU clock frequency and the input supply voltage for Programmable Logic (PL) unit on an FPGA-based designed processing unit. Experiments were performed on the ZedBoard device that includes Zynq 7000 AP - SoC. Results show that the DVFS technique is more effective on dynamic power consumption compared to static power which is changing when the input power supply Vccint reaches the maximum value. Also, the value of dynamic power consumption appears to be more affected than the static power consumption. When the frequency is increased and Vccint is constant, static power consumption is not changing so much in value, while the dynamic power consumption is improved by an additional 3%.

The contributions of this work can be summarized as follows:

- We introduce the implementation of the DVFS technique for low power consumption, on a designed processing unit on a hard IP Core Zynq 7000 SoC.
- This work analyses the power consumption and the scaling capabilities of high-density 28 nm Zynq 7000 AP SoC element from the perspective of the hardware designer concerning IP cores.

As future work we will focus on the design of processing unit architectures that will be used for a real-time application, to evaluate power consumption. Another research point in this area is the implementation of an optimization algorithm for Zynq 7000 SoC-based systems.

ACKNOWLEDGEMENT

The results in this paper are partly supported by the financing of mobilities in the CEEPUS Network: CIII-BG-1103-04-1920-Modelling, Simulation and Computer-aided Design in Engineering and Management.

REFERENCES

[1] Kuon, I., Rose, J.: 'Measuring the gap between FPGAs and ASICs', IEEE Trans. Computer-Aided Design Integrated Circuits Systems, 2007, 26, (2), pp. 203–215

[2] C. T. Chow, L. S. M. Tsui, P. H. W. Leong, W. Luk, and S. J. E. Wilton, "Dynamic voltage scaling for commercial FPGAs," in Proceedings of International Conference on Field-Programmable Technology (FPT), pp. 173–180, Dec 2005.

[3] UG907 - Vivado Design Suite User Guide Power Analysis and Optimization (v2019.2) October 30, 2019

[4] Rahman, A., Das, S., Tuan, T., Rahut, A.: 'Heterogeneous routing architecture for low-power FPGA fabric'. Proc. of the IEEE Custom Integrated Circuits Conf., Sept. 2005, pp. 183–186

[5] Ryan, J., Calhoun, B.: 'A sub-threshold FPGA with low-swing dual-Vdd interconnect in 90 nm CMOS'. Proc. 2010 IEEE Custom Integrated Circuits Conf. (CICC), 2010, pp. 1–4

[6] Li, F., Lin, Y., He, L.: 'Vdd programmability to reduce FPGA interconnect power'. IEEE/ACM Int. Conf. on Computer-Aided Design, 2004 (ICCAD-2004), 2004, pp. 760–765

[7] Li, F., Lin, Y., He, L., Cong, J.: 'Low-power FPGA using pre-defined dual-Vdd/dual-vt fabrics'. Proc. 2004 ACM/SIGDA 12th Int. Symp. on Field Programmable Gate Arrays FPGA '04. ACM, New York, NY, USA, 2004, pp. 42–50

[8] Raham, A., Polavarapuv, V.: 'Evaluation of low leakage design techniques for field-programmable gate arrays'. Proc. 2004 ACM/ SIGDA 12th Int. Symp. on Field Programmable Gate Arrays (FPGA '04) ACM, New York, NY, USA, 2004, pp. 23–30

[9] Lamoureux, J., Wilton, S.: 'On the interaction between power-aware FPGA cad algorithms. Int. Conf. on. Computer-Aided Design (ICCAD-2003), 2003, pp. 701–708
[10] Lamoureux, J., Wilton, S.: 'Clock-aware placement for FPGAs'. Int. Conf. on Field Programmable Logic and Applications, 2007 (FPL 2007), 2007, pp. 124–131
[11] Gayasen, A., Tsai, Y., Vijaykrishnan, N., Kandemir, M., Irwin, M.J., Tuan, T.: 'Reducing leakage energy in FPGAs using region constrained placement'. Proc. 2004 ACM/SIGDA 12th Int. Symp. on Field Programmable Gate Arrays (FPGA '04) ACM, New York, NY, USA, 2004, pp. 51–58
[12] Nabina, A., Nunez-Yanez, J.: 'Adaptive voltage scaling in a dynamically reconfigurable FPGA-based platform'. ACM Trans. Reconfigurable Technol. Syst. 5, 4, Article 20 (December 2012)
[13] UG953 Vivado Design Suite 7 Series FPGA and Zynq-7000 SoC Libraries Guide
[14] "Zynq-7000 SoC Technical Reference Manual UG585 (V1.12.2)." Technical Reference Manual, Zynq-7000 All Programmable SoC, Xilinx, Inc., July 2018.
[15] M. Sadri, C. Weis, N. When, and L. Benini, "Energy and Performance Exploration of Accelerator Coherency Port Using Xilinx ZYNQ", Proc. 10th FPGA World Conference, Copenhagen/Stockholm, 2013.
[16] A.K. Jain, K.D. Pham, J. Cui, S.A. Fahmy, and D.L. Maskell, "Virtualized Execution and Management of Hardware Tasks on a Hybrid ARM-FPGA Platform", Journal of Signal Processing Systems, vol. 77, no. 1-2, pp. 61-76, 2014
[17] Xilinx, Zynq-7000 AP SoC Low Power Techniques part 3 - Measuring ZC702 Power with a Standalone Application Tech Tip, 2014.
[18] Beldachi, A. F., & Nunez-Yanez, J. L. (2014). Accurate power control and monitoring in ZYNQ boards. 2014 24th International Conference on Field Programmable Logic and Applications (FPL).

Design of a sampling mixer for use in UWB radar applications

Marko Malajner
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
marko.malajner@um.si

Dušan Gleich
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
dusan.gleich@um.si

***Abstract*—The design, simulations and measurements of a sampling mixer are presented in this paper. Ultra-Wide-Band pulse radar applications demand high performance digitalization of the received RF signal with few GHz of bandwidth. Since high sampling rate analog to digital converters are very expensive and complex, we designed an inexpensive sampling mixer based on the equivalent sample time technique. A sampling mixer uses a bridge with two diodes for capturing a portion of a signal in each period. After a certain amount of repetitive periods, the sampling mixer reconstructs the original RF signal. The reconstructed signal is stretched from the sub-nanoseconds order into the microsecond order. Such stretched RF signal can be digitalized using a low rate analog to digital converter with sampling rate of a few MSa/s.**

***Index Terms*—pulse UWB radar, sampling mixer, equivalent sample time, ground penetrating radar**

I. Introduction

Ultra-wideband (UWB) radars produce very short Radio-Frequency (RF) pulses in the sub-nanoseconds order for sensing and imaging application. UWB pulses have good spatial resolution and penetration in dielectric materials. On the receiver side, UWB radar measures the reflected signals that arise due to the difference in the electrical properties between the observed object and surrounding environment [1]. Radars in the UWB domain have a wide area of uses, from military applications, like buried mine detection, to commercial applications like medical imaging, ground penetration radars, etc. [2], [3]. The principal UWB radar consist of a transmitter with a pulse generator, a receiver with pulse detector, antennas and a signal processing unit [4]. On the transmitter side, the circuit must be capable of producing sub-nanosecond pulses [5]. The second important and more complicated part of UWB radar is the receiver. The receiver must receive and digitalize wideband signals in the order of a few GHz. One could use a high sample-rate Analog to Digital Converter (ADC) for direct digitalization of received signals, and such ADC must have sample rates above few GSa/s. Texas Instruments [6] offers an ADC up to 10 GSa/s, however, high sampling-rate ADCs are too expensive for using in UWB radars, and produce large amounts of data in the orders of tens and more GB per time unit [7]. To overcome the drawbacks of direct AD sampling, the Equivalent Sample Time (EST) samplers are used instead of high sample-rate ADCs [8]. This kind of samplers captures a short part of the signal in each period, and then combines them to reconstruct the original signal. EST samplers cannot capture a whole signal in one shot, and therefore needs many repetitions. With this technique, the sampler stretches the original signal from the sub-nanosecond order to the micro or milliseconds order. A stretched signal is easy to digitalize with inexpensive commonly used ADCs.

In this paper, we present the simulation, design and development of a improved sampling mixer which is able to capture signals in the sub-nanosecond order and convert them in–to the microsecond order. The improved strobe generator used in the sampling mixer made the sampling mixer much simpler. The used commercially available balun for splitting the strobe signal in the improved sampling mixer has a smaller signal loss in comparison to a radial slot stub balun. Such stretched signal is then straightforward to digitalize using low sample rate AD Converters.

The rest of the paper is organized as follows. Section II describes the design of the sampling mixer, from simulation to design of the Printed Circuit Board (PCB). Section III describes the measuring of sampler performance, and the last Section concludes the paper.

II. Development of the sampling mixer

Generating of sub-nanoseconds' pulses is the easy part of development in comparison to development of an RF receiver. The 120 picosecond pulse, generated with Step Recovery Diodes (SRD), has a bandwidth around 8 GHz, which means that the receiver must be capable of converting an 8 GHz bandwidth analog signal to digital. To satisfy the Shannon theorem, the ADC must have sample-rate at least twice the frequency of the input RF signal. Analog to Digital Converters with tens of GSa/s are extremely expensive, and, therefore, unsuitable for use in low cost radars. The solution is to use an equivalent sample time ADC. A realistic approach is to sub-sample the RF signal upon extending its time scale from picosecond into nanoseconds. Thus, the extended time scale of an RF signal can be handled by conventional ADCs. In literature, the sampling mixer is also called a sampling head or signal stretcher. In the rest of paper we will simply name it the sampler. There are many different sampling concepts, and we chose and redesigned the sampler reported in [9].

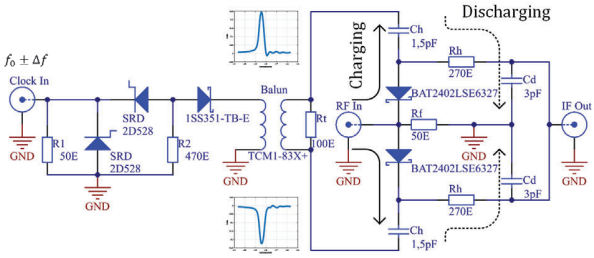


Fig. 1: Schematic of the proposed sampling mixer. Left from the balun is the pulse generator. After the balun, which split the pulse from the generator into opposite pulses, is a sampling bridge with two diodes. The graphs present measured pulses at the balun's output.

The basic principal of sampling is the repeated quasi-instantaneous capturing of a time-varying signal by a sampling gate. The gate is opened and closed by a time precise train of narrow strobe pulses. During sampling, the repetitive scattered pulses from the receiving antenna are applied to the gate with Pulse Repetition Frequency (PRF) f_0. Strobe pulses on the sampling gate are triggered with the slightly offset frequency given by $f_0 \pm \Delta f$. The received RF signal and strobe signal are mixed in a way that the strobe signal scans across the sampled RF signal. The speed of a complete sampling cycle is defined by ${}^{1}/_{\Delta f}$. An extending factor α is defined as ${}^{f_0}/_{\Delta f}$, and is equal to an extending ratio between the original received signal and the reconstructed signal.

The sampler is based on a two-diode bridge. Fig. 1 shows the schematic of the developed sampler. The values of components are based on simulation results using an ADS simulation tool [10]. Also, the whole PCB was designed and simulated on a simulation tool. The diodes in the bridge are low capacitance ($C = 0.2pF$) RF Schottky. Its low barrier height, small forward voltage and low junction capacitance make BAT24-02LS a suitable choice for mixer and detector functions in applications in which frequencies are as high as 24 GHz [11].

The main challenge in the construction of a sampler is to achieve very short strobe opposite pulses with high amplitudes. The amplitude of pulse must be able to open the bridge diode, in our case at least 0.25 V. For generating pulses, we used an improved two SR diodes generator, reported in [12]. The reported generator used an additional negative supply voltage. Instead of this negative supply voltage, we used a precise clock generator with symmetric ±0,5 V amplitude and 500 ps rise time. The PIN diode on input is also removed due to this simplification. The PIN diode in the original research was used for blocking negative input voltage. SR diodes were biased with an adjustable negative supply voltage in order to push SRDs into a reverse regime. The SRD generator, triggered with an amplified clock signal [13], produced a pulse with amplitude of approximately 1.2 V. This pulse is then divided into two pulses with opposite polarities using the balun. We chose an SMD fabricated balun manufactured by the Mini-Circuit company instead of a radial slot stub balun, due to less signal loss. A positive pulse opens the upper diode and a negative pulse opens the lower diode, according to Fig. 1. The TCM1-83X+ balun has a wide frequency range from 10 to 8000 MHz [14]. Measured pulses on the balun output are shown in Fig. 2. At around 0.3 V, where the barrier of the diode is, the pulse duration is about 100 ps (measured with an oscilloscope). In other words, the sampler bridge is capturing around 100 ps of the RF input signal in one repetition. This part of the energy is stored in capacitor C_d.

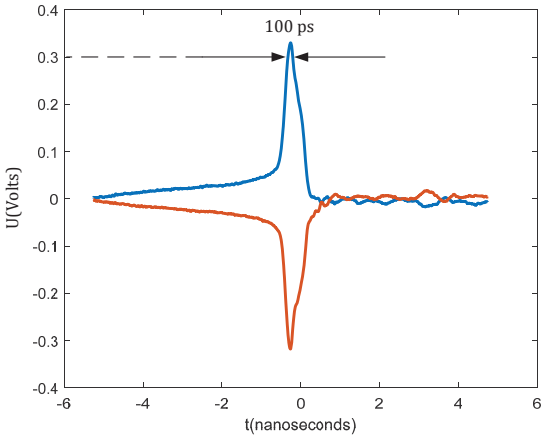


Fig. 2: Measured strobe 100 ps pulse duration at diode forward voltage.

The captured RF signal is stored in a charging capacitor C_h, and, together with the series resistance of the diode ($R_s = 8Ohms$), provides an RF charging time around 12 ps. Resistor R_h and capacitor C_d provide the discharging RC network with a time constant much slower than the time constant of the charging network. Values of C_d and R_h were determined using the simulation tool in order to obtain good signal conversion loss.

A. PCB design

The next step after successful simulation of the sampler was development of a Printed Circuit Board (PCB). When designing an RF PCB, the designer must take into account many parameters, such as line width and length, thickness of the board, dielectric constant of the material, shape of the corners, etc. To produce such PCB takes many iterations without using a simulation tool. We used the ADS simulation tool from Keysight [10] in order to proceed with simulation of the whole outline of the PCB. Fig. 3 shows the designed and simulated outline of the sampler.

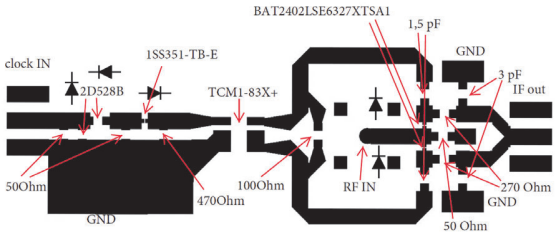
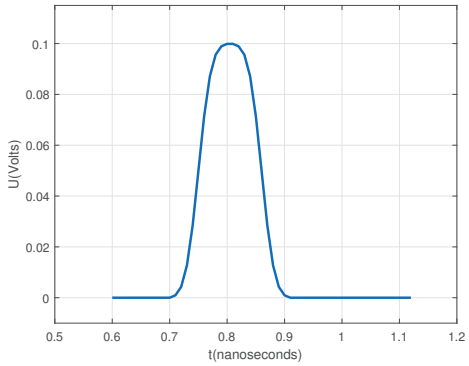
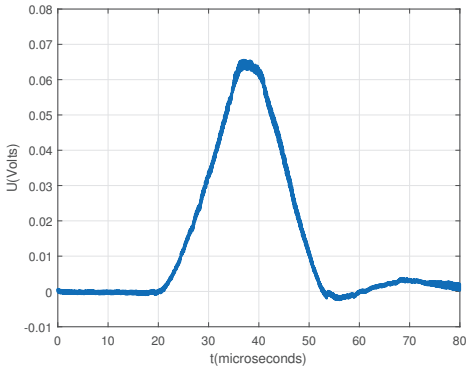


Fig. 3: Simulated outline of sampler PCB

In the simulation we simulated a 200 ps width pulse at the input RF port (denoted as RF IN in Fig. 3). The input pulse had a peak voltage of 0.1 V and rise/fall time of 0.1 ns. Strobe frequency f_0 was set to 10 MHz and frequency Δf to 100 Hz. The sampler stretched the input pulse from 200 ps to 20 μs, which confirmed extending factor α, defined as ${}^{f_0}/_{\Delta f} = {}^{10MHz}/_{100Hz} = 100000$.



(a) 200 ps RF input pulse



(b) 20 μs stretched pulse at output

Fig. 4: Simulation result of sampler

III. Measuring Results

The PCB was fabricated on two sided ROGERS RO4350B laminate with the dielectric constant 3.66 and substrate thickness 0.762 mm. The fabricated sampler, pulse generator and clock source are depicted in Fig. 5. After fabrication, we took measurements of the sampler. Instead of using amplifiers and antennas, we connected the pulse generator output directly via an attenuator with the RF sampler input. In such a way, the scattering signals from the environment are eliminated if the measurements are not conducted in an anechoic chamber. A synchronized 2 output channels clock CDC6208 was used for generating pulses and for generating the strobe signal for sampling. Measurements are shown in Fig. 6. Generated RF pulses were measured using an Agilent 40 GSa/s oscilloscope, meanwhile the IF output of the sampler was measured using a Rigol 2 GSa/s oscilloscope. We can observe in Fig. 6 that the sampled pulse was stretched 100,000 times, and that measurements confirmed the simulation results. The IF output of the sampler can now be digitalized using an ADC with a 5 MSa/s or less sampling rate.

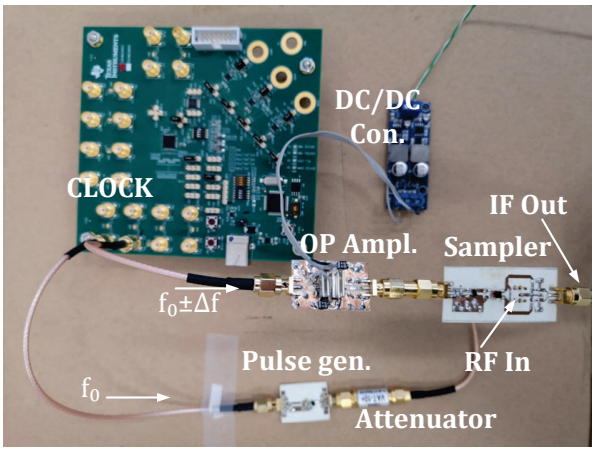
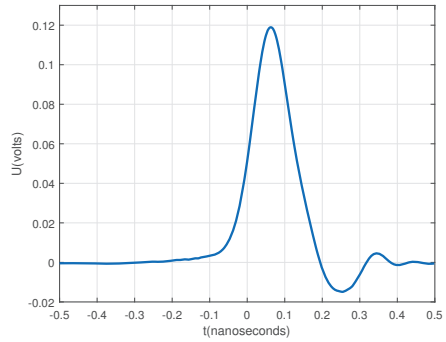


Fig. 5: Prototype of radar without antennas ready for measurements

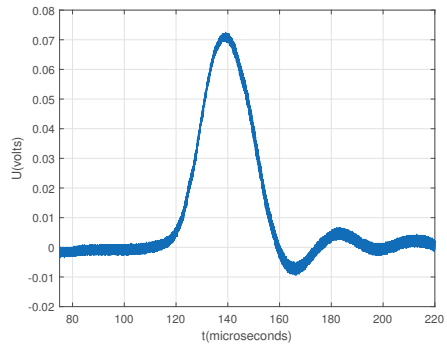
A comparison between oscilloscope signal and sampler signal was made to validate the sampler signal acquisition. In Fig. 7 are both signals (from oscilloscope and sampler), normalized and scaled. From the signals were calculated relative error using equation: $e = \frac{signal_{sampler} - signal_{ocilloscope}}{signal_{ocilloscope}}$ The yellow curve in Fig. 7 represents relative error. The peak value of relative error is below 15 %, and average relative error is below 1 %. In addition, a cross-correlation was made between pulses obtained by the oscilloscope and sampling mixer. Cross-correlation is depicted in Fig. 8.

IV. Conclusion

In this paper, we presented the simulation and design of a low cost sampling mixer which is capable of sampling a signal at the rate of 100 GSa/s, calculated in equivalent sampling time. The improved sampling mixer consists of commercially available components, and it was fabricated on Rogers laminate. The sampling mixer could be assembled in any laboratory without using expensive tools and equipment.



(a) Generated RF pulse



(b) Sampled pulse

Fig. 6: Measured results of sampler

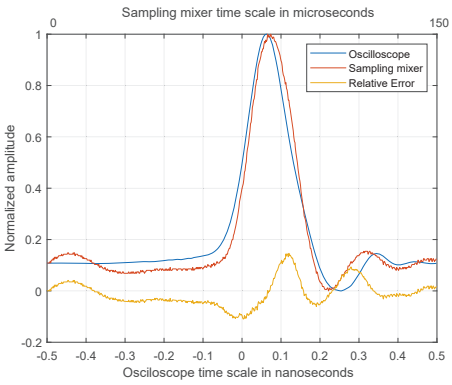


Fig. 7: Comparison of signal acquisition between oscilloscope and sampler. The yellow curve is relative error.

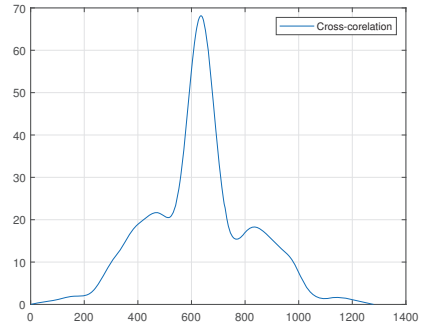


Fig. 8: Cross-correlation between the oscilloscope sampled and sampling mixer sampled RF pulse

Measurements of sampler performance were confirmed by the simulation results. In comparison with a commercial ADC, the sampling mixer is more than 10 times cheaper.

We showed that the sampling mixer is capable of stretching a signal from the hundred picosecond in to the microsecond order. Such sampler has potential in many radar applications where high rate sampling is needed.

References

[1] L. Li, A. E. Tan, K. Jhamb, and K. Rambabu, "Buried object characterization using ultra-wideband ground penetrating radar," *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, pp. 2654–2664, Aug 2012.

[2] D. Oloumi, J. Ting, and K. Rambabu, "Design of pulse characteristics for near-field uwb-sar imaging," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, pp. 2684–2693, Aug 2016.

[3] R. Feghhi, D. Oloumi, and K. Rambabu, "Design and development of an inexpensive sub-nanosecond gaussian pulse transmitter," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, pp. 3773–3782, Sep. 2019.

[4] L. Liu, X. Xia, S. Ye, J. Shao, and G. Fang, "Development of a novel, compact, balanced, micropower impulse radar for nondestructive applications," *IEEE Sensors Journal*, vol. 15, pp. 855–863, Feb 2015.

[5] D. Šipoš, M. Malajner, and D. Gleich, "Steeped frequency and uwb pulse based radars for landmine detection," in *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 27–30, June 2019.

[6] "Analog-to-digital converters (adcs) — products — data converters — ti.com," 2020.

[7] A. S. Venkatachalam, X. Xu, D. Huston, and T. Xia, "Development of a new high speed dual-channel impulse ground penetrating radar," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, pp. 753–760, March 2014.

[8] J. Han and C. Nguyen, "Development of a tunable multiband uwb radar sensor and its applications to subsurface sensing," *IEEE Sensors Journal*, vol. 7, pp. 51–58, Jan 2007.

[9] C. Zhang, A. E. Fathy, and M. Mahfouz, "Performance enhancement of a sub-sampling circuit for ultra-wideband signal processing," *IEEE Microwave and Wireless Components Letters*, vol. 17, pp. 873–875, Dec 2007.

[10] "Pathwave advanced design system (ads) — keysight," 2020.

[11] "Bat24-02ls - infineon technologies," 2020.

[12] L. Zou, S. Gupta, and C. Caloz, "A simple picosecond pulse generator based on a pair of step recovery diodes," *IEEE Microwave and Wireless Components Letters*, vol. 27, pp. 467–469, May 2017.

[13] "Cdcm6208 2:8 ultra low power, low jitter clock generator — ti.com," 2020.

[14] "Mini-circuits," 2020.

Interference classification for IEEE 802.15.4 networks

Uros Pesovic
Faculty of technical sciences Cacak
University of Kragujevac
Cacak, Serbia
uros.pesovic @ftn.kg.ac.rs

Sladjana Djurasevic
Faculty of technical sciences Cacak
University of Kragujevac
Cacak, Serbia
sladjana.djurasevic@ftn.kg.ac.rs

Vanja Lukovic
Faculty of technical sciences Cacak
University of Kragujevac
Cacak, Serbia
vanja.lukovic@ftn.kg.ac.rs

Peter Planinsic
Faculty of electrical engineering and computer science
University of Maribor
Maribor, Slovenia
peter.planinsic@um.si

***Abstract*—Wireless sensor networks most commonly operate in 2.4 GHz band using the IEEE 802.15.4 standard for wireless transmission between sensor nodes. This frequency band is also used by IEEE 802.11 and IEEE 802.15.1 networks which operate with higher transmitting powers and could cause significant inter-protocol interference, especially in highly urbanized areas. To improve coexistence in 2.4 GHz band, IEEE 802.15.4 networks must to be able to identify and evade these interferences. In this paper, we present practical implementation of interference classification algorithm based on the k-means clustering which could be used by low-cost commercially available IEEE 802.15.4 transceivers.**

Keywords—IEEE 802.15.4, coexistence, interference classification, k-means clustering, machine learning

I. INTRODUCTION

Wireless sensor networks are used for remote sensing of various environmental parameters on a broad area of interest. They employ a large number of sensor nodes to collect data from their sensors and transmit data wirelessly towards the central node for further processing. Radio transmission is the biggest energy consumer in wireless sensor nodes which is battery-powered, so nodes need to operate efficiently to achieve months or years of operation with a single set of batteries. Wireless sensor nodes communicate most often in 2.4GHz ISM band using low-power transceivers compliant with IEEE 802.15.4 standard. This band is license-free worldwide and is also used by other wireless networks, operating under IEEE 802.1b/g/n standard and IEEE 802.15.1 standard a shown in Fig 1. Due to their higher transmitting powers and higher-data rates, these networks causes significant inter-protocol interference to IEEE 802.15.4 devices, especially in densely populated urban areas. Thus IEEE 802.15.4 devices need to implement various mitigation and interference avoidance strategies to coexist with these networks in the same 2.4 GHz band [1].

IEEE 802.15.4 standard divides 2.4 GHz band into 16 channels, numbered from 11 to 26. Each channel has 2 MHz bandwidth, with 5 MHz separation between adjacent channels. On the other hand, IEEE 802.11 standard in 2.4 GHz band, defines 11 channels for the United States and 13 channels for Europe. These channels have 20 MHz bandwidth (22 MHz for IEEE 802.11b) and partially overlap with 5 MHz separation between channel center frequencies. Thus one IEEE 802.11b/g channel will interfere with four IEEE 802.15.4 channels. In the case of the IEEE 802.11n standard which uses MIMO-OFDM, two non-overlapping IEEE 802.11 channels were used to form 40 MHz wide channels which could overlap with up to eight IEEE 802.15.4 channels. Due to the high transmitting power of IEEE 802.11 devices, set to a maximum of 100 mW, these networks will significantly interfere in the operation of much weaker IEEE 802.15.4 devices which have maximum transmitting power of 1 mW. IEEE 802.11 types of networks are frequently used for local area networking in homes and apartments since it does not require any infrastructure and most devices such as laptops and smartphones support this standard. Due to the increased number of wireless devices and the increase of usage of internet services in homes, these kinds of networks utilize much of the 2.4GHz band, especially in highly urbanized areas. This poses significant problems for IEEE 802.15.4 devices which are in such areas used for smart home automation, smart metering, etc. Newly adopted IEEE 802.11ac standards could solve this issue in the future by using higher frequency bands for Gigabit local area networking, thus significantly reducing the utilization of 2.4 GHz ISM band.

IEEE 802.15.1 standard known as Bluetooth divides 2.4 GHz band into 79 channels with 1 MHz bandwidth. Bluetooth uses a frequency-hopping spread spectrum and performs 1600 channel hops per second. A newer version of Bluetooth Low Energy standard divide this band intro 40 channels with 2 MHz bandwidth, and uses adaptive frequency hopping to avoid congested channels. Due to frequent channel hopping and pseudorandom hopping pattern, interference from IEEE 802.15.1 devices is difficult for IEEE 802.15.4 to detect and to mitigate this kind of interference. Transmitting power of IEEE 802.15.1 devices ranges depending on device class from 0.5 mW (Class 4) to a maximum of 100 mW (Class 1). Bluetooth is most commonly used for wireless audio streaming using headsets and speakers. Due to lower data-rates, channel hopping and sparse utilization in homes and apartments, Bluetooth causes

much less interference to IEEE 802.15.4 devices compared to IEEE 802.11 networks.

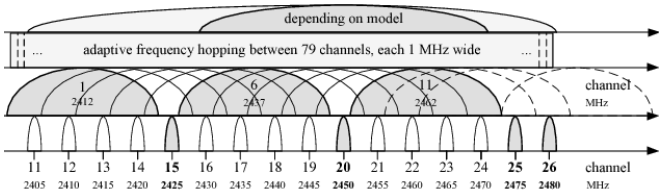


Fig. 1. Channel division of 2.4 GHz band for various wireless standards

The coexistence of IEEE 802.15.4 networks in highly utilized 2.4 GHz band can be improved using proactive and reactive techniques. Proactive techniques relay on coordinated effort in planning and deployment of various types of networks in a certain area. This can be achieved by spatial separation of networks, frequency separation or time separation. This approach requires control over deployed networks which is typically not the case in urban areas. In such a case, reactive techniques rely on the ability of network coordinator to detect and mitigate interference without the need to make any changes to the interfering network. Reactive measures either suspend IEEE 802.15.4 network operation for a certain time until the interference is present in the channel or change the operating channel of the entire network. Certain solutions propose a time-slotted channel hopping technique, which performs channel hopping to avoid congested channels [2-5].

For that reason, IEEE 802.15.4 devices need to have the ability to identify and classify interference in order to make the right decisions during the interference mitigation process. These devices use CSMA/CA (Carrier Sense Medium Access/Collision Avoidance) which first checks the channel state using Clear Channel Assignment (CCA) before starting the transmission of a packet. CCA is performed for 8 symbol periods (128 μs) and can be performed either by Carrier Sense (CS) which detects IEEE 802.15.4 carrier signal or by performing Energy Detection (ED) in the channel. Carrier sense CCA is intended to detect transmission of other IEEE 802.15.4 devices, while ED can be used to detect transmission of other types of networks that share the same frequency and it will be reported in form of RSSI (Received Signal Strength Indicator). ED is also used by MAC during the initial formation of the network to estimate the state of all 16 channels and choose the channel with the lowest energy level. ED can be used by IEEE 802.15.4 network coordinator to perform a sweeping scan of all available channels in order to measure how much interference is present in each channel. Radio transmission from the same IEEE 802.11 interferer will occupy several IEEE 802.15.4 channels. Since IEEE 802.11g/n power spectrum is flat for the entire IEEE 802.11 channel bandwidth (Fig 2.), ED measurements in IEEE 802.15.4 channels which are overlapped with this type of interference will have similar RSSI values [6]. Thus these values can be clustered together to a certain point which could be identified as an interferer.

II. Interference Classification

Clustering is a method of multivariate analysis used to classify observations (instances) so that instances within a group are similar to each other and significantly different between objects within other groups.

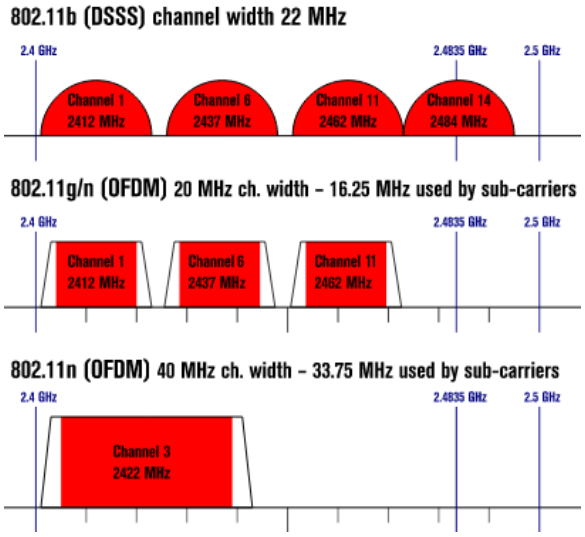


Fig. 2. Spectrum of non-overlapping IEEE 802.11b/g/n channels

The main purpose of clustering is to find a "natural" grouping of a set of instances based on the different characteristics of each instance. The most commonly used algorithm for clustering is known as k-means clustering. K-means algorithm is based on spherical clusters that are separated in such a way that the value of mean converges towards the center of the cluster, as shown in Fig 3. K-means algorithm starts with random placement of k-cluster centroids and assigning instances to the closest cluster centroid. Distances are most commonly measured using Euclidian distance, but also Manhattan, Minkowski or Chebychev distance metric could be used as well. For the newly formed cluster, attributes of all instances within the cluster are averaged to find the new position of cluster centroid. This process is repeated until cluster centroids converge or after a certain number of iterations. Effectiveness of clustering can be measured by centroid separation or using cost function which calculated the average standard deviation of instances from their assigned cluster centroids.

The result of k-means clustering depends on the initial position of k cluster centroid points. Since these points are randomly chosen, depending on their initial position, an algorithm can converge slower or faster, or in certain cases can end up at a local minimum. Initialization can be improved by “forgy” algorithm which instead of random centroid points chooses k random instances for centroid points or by Kaufman algorithm which uses most centralized instances for centroid points. These initialization algorithms could significantly speed up the convergence process. To avoid convergence in a local minimum algorithm is repeated several times and cluster centroid points which have minimal cost function is chosen as an optimal solution.

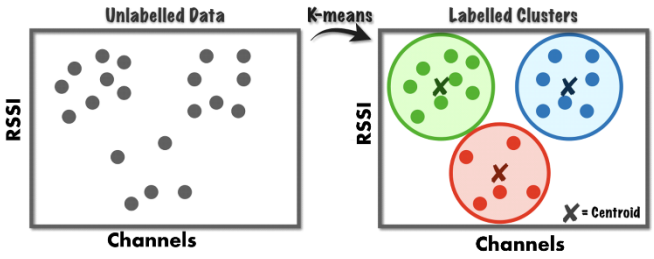


Fig. 3. K-means algorithm for interference clustering

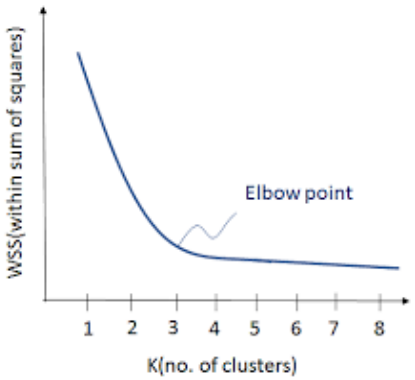


Fig. 4. Elbow method for determining number of clusters

A challenging task in clustering is determining the sufficient number of clusters. Since clustering is a branch of unsupervised learning, there is no previous knowledge about the expected number of clusters. This can be determined iteratively using the elbow method, starting with the small number of clusters and then incrementing the number of clusters until cost functions reach the "elbow" point, which represents an optimal number of clusters as shown on Fig. 4 . The optimal number of a cluster can be also determined using the Average silhouette or Gap statistic method.

III. EXPERIMENTAL RESULTS

IEEE 802.15.4 channels in 2.4 GHz ISM band are observed using the PICDEM Z development board [7]. This board uses MRF24J40MA [8], 2.4 GHz IEEE 802.15.4 compliant transceiver, realized on prefabricated Printed Circuit Board (PCB) with a dipole antenna. It communicates with a transceiver using a four-wire Serial Peripheral Interface (SPI). MRF24J40 transceiver supports ZigBee, MiWi, MiWi P2P and Proprietary wireless networking protocols. It operated in all sixteen channels in ISM Band 2.405-2.48 GHz in data rates: 250 kbps (IEEE 802.15.4) and 625 kbps (proprietary turbo mode). The transmitter has 0 dBm typical output power with 36 dB power control range while the receiver has typical sensitivity of -95 dBm. This transceiver can perform ED for a custom number of symbols, which range from 1,2, 4 or 8 symbols. The averaged value of RSSI represents the estimate of the received signal power within the bandwidth of an IEEE 802.15.4 channel. The RSSI value is an 8-bit value ranging from 0-255 where the mapping between the RSSI values with the received power level is shown in Fig. 5.

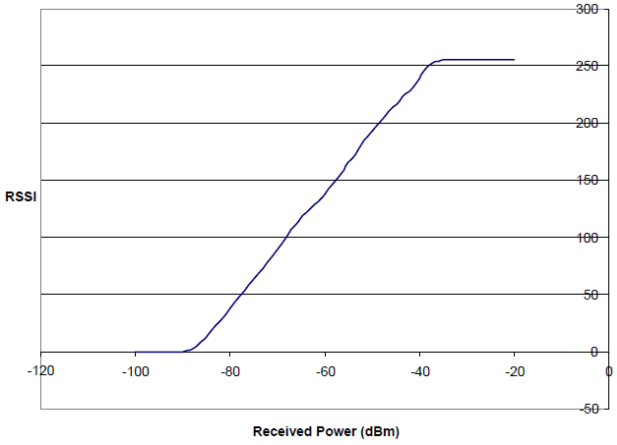


Fig. 5. RSSI mapping to received power level

During the interference classification, the microcontroller puts the MRF24J40MA transceiver is put into reception mode in which performs a cyclic sweeping scan of all channels as shown in Fig6. For each channel, an ED scan is performed for 8 symbol periods (128 µs) as required by the IEEE 802.15.4 standard. ED is issued by setting RSSI start bit, which initiates RSSI averaging, after which RRSI ready bit is set and results can be read from RSSINUM register.

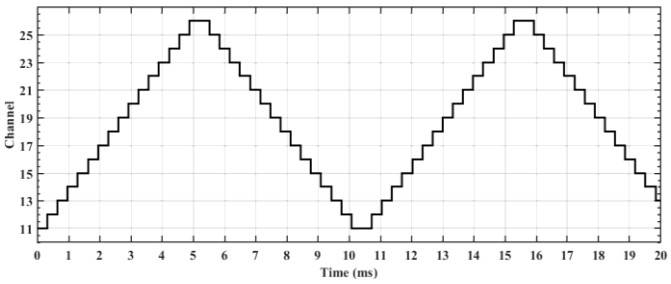


Fig. 6. Channel sweep pattern during ED scan

When ED in the current channel is finished, the adjacent channel is selected by setting a channel number in the RFCON register. Transition to a new channel requires 192 µs delay to enable sufficient time for stabilization of transceiver oscillator for new center frequency. This time is used to transmit channel number and RSSI value to host computer via UART at a transmission speed of 115200 bps. Total scan time per one channel is 320 µs or 5.12 ms to scan all sixteen channels.

RSSI results, collected by the host computer for multiple channel scans are logged and then processed in Matlab. First, the background noise floor is determined, and all RSSI instances which are below this value are not used in the clustering algorithm. As shown in Fig 7, the result of the k-means clustering algorithm located two clusters of RSSI values. According to the standard deviation of instances within the cluster, we can classify that cluster occupies four IEEE 802.15.4 channels which implies that this type of interference is originating from IEEE 802.11g wireless networks. In the case of IEEE 802.11n networks, up to eight neighboring channels will be occupied. Our goal is to implement a clustering algorithm on a microcontroller that has very limited computational resources. Obtained information from this algorithm can be used by the IEEE 802.15.4 network coordinator to avoid operation in channels that are occupied by classified interferences thus improving IEEE 802.15.4 coexistence.

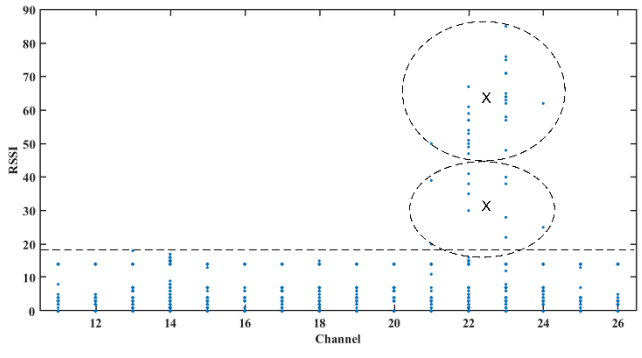


Fig. 7. Clustering of two IEEE 802.11 interferences

IV. Conclusion

This work presented a method for interference classification using the k-means clustering algorithm applied on RSSI data collected by ED scans of all IEEE 802.15.4 channels. Received RSSI data is clustered together which by statistical analysis of cluster itself enables classification of interference type. This approach enables these kinds of networks to improve coexistence in case of interference of IEEE 802.11 networks which are frequently present in urbanized environments, such as public and residential buildings. Further work will be focused on the implementation of clustering algorithms for interference classification, which is suitable for implementation on systems with constrained resources such are wireless sensor nodes.

Acknowledgment

This study was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and these results are parts of the Grant No. 451-03-68/2020-14/200132 with University of Kragujevac - Faculty of Technical Sciences Čačak.

References

[1] Co-existence of IEEE 802.15.4 at 2.4 GHz", NXP Application Note JN-AN-1079, Revision 1.1, 8-Nov-2013

[2] S. Grimaldi, A. Mahmood, M. Gidlund, "Real-Time Interference Identification via Supervised Learning: Embedding Coexistence Awareness in IoT Devices", IEEE Access (Volume: 7) 2018, pp: 835 - 850, ISSN: 2169-3536

[3] H. Dakdouk, E. Tarazona, R. Alami, R. Feraud, G. Z. Papadopoulos, P. Maille, "Reinforcement Learning Techniques for Optimized Channel Hopping in IEEE 802.15.4-TSCH Networks", MSWIM '18 Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 99-107, Montreal, Canada, 28.10 - 02.11, 2018

[4] S. Hammoudi, S. Harous, Z. Aliouat and L. Louail, "Time slotted channel hopping with collision avoidance", Int. Journal Ad Hoc and Ubiquitous Computing, Vol. 29, Nos. 1/2, 2018 85

[5] A. Elsts, X. Fafoutis, R. Piechocki, and I. Craddock, Adaptive Channel Selection in IEEE 802.15.4 TSCH Networks, 2017 Global Internet of Things Summit (GIoTS), 6-9 June 2017 Geneva, Switzerland, ISBN: 978-1-5090-5873-0

[6] N. H. Mahalin, H. S. Sharifah , S.K. Syed Yusof, N. Fisal, R.A Rashid, "RSSI Measurements for Enabling IEEE 802.15.4 Coexistence with IEEE 802.11b/g", TENCON 2009 - 2009 IEEE Region 10 Conference, 23-26 Jan. 2009, Singapore, ISSN: 2159-3450

[7] PICDEM Z Demonstration Kit User's Guide, Microchip Technology Incorporated,2008,

[8] Microchip, "MRF24J40MA Datasheet, 2.4 GHz IEEE Std. 802.15.4™RF Transceiver Module", Microchip Technology Incorporated,2008

USRP Implementation of a Ground Penetrating Radar Using a Combination of Stepped Frequency and OFDM Principles

Venceslav Kafedziski, Sinisha Pecov, Dimitar Tanevski
Faculty of Electrical Engineering and Information Technologies
University Ss Cyril and Methodius
Skopje, Republic of North Macedonia
{kafedzi@feit.ukim.edu.mk, nine_pec@yahoo.com, dimitar-tanevski@hotmail.com}

***Abstract*—We present a Software Defined Radio (SDR) implementation of a Ground Penetrating Radar which uses much larger bandwidth than the SDR instantaneous RF bandwidth in order to increase radar resolution. It is based on the Stepped Frequency principle, where the baseband signal is an Orthogonal Frequency Division Multiplexing (OFDM) signal, transmitted in different subbands by modulating the RF stepped frequencies. Using OFDM in each subband provides faster signal transmission and simpler detection. The receive signal-to-noise ratio (SNR) is improved by repeating a number of times the transmit baseband signal before modulating each RF frequency. The random phase discontinuities between the adjacent subbands that appear when the RF frequency is changed, are removed by postprocessing the received signal. Singular value decomposition is used to remove the effects of the direct and the ground reflected waves. We perform field experiments with anti-tank mines buried in the ground. The resulting B-scans provide excellent detection and localization of the buried objects.**

***Index Terms*—Ground Penetrating Radar, Stepped Frequency Radar, OFDM Radar, Software Defined Radio, USRP**

I. Introduction

Software-defined radio (SDR) is a radio communication system where components that have been traditionally implemented in hardware (e.g. mixers, filters, amplifiers, modulators/demodulators, detectors, etc.) are instead implemented by means of software on a personal computer or embedded system. This approach provides lightweight and cheap implementation of different communication systems, especially for prototyping and testing purposes. However, as the SDR is not designed for the dedicated purpose, it might not meet some of the requirements of the specific system. Our goal is to design a Ground Penetrating Radar (GPR) using an SDR device. GPR is a nondestructive method that emits electromagnetic waves into the ground, and records the reflected/scattered signals from subsurface objects and structures, in order to determine their location, and, possibly, their shape. GPR is also used to study bedrock, soils, groundwater, and ice. It is studied in many references, e.g. [1]. There is an extensive literature on the subject of implementing an SDR radar, for use in the air or for locating objects under the ground surface, i.e GPR. Most of the implementations use different versions of the Universal Software Radio Peripheral (USRP) of the company Ettus, owned by National Instruments. SDR radars use either Frequency Modulation, i.e. the Frequency Modulation Continuous Wave (FMCW) radars, or a set of pulses with different frequencies that are sent subsequently, i.e. the Stepped Frequency (SF) radars. Orthogonal Frequency Division Multiplexing (OFDM) radars that send all the pulses with different frequencies simultaneously, are also used. It is well known that the radar resolution depends on the bandwidth used. Most of the SDR implementations use the instantaneous SDR bandwidth, and, thus, have low resolution, which is insufficient for detecting small objects. Early works in this area are [2], where an approach to USRP GPR is described in principle using USRP WBX, but tested with only a sinusoidal signal, and [3], where NI-USRP 2920 FMCW/OFDM radar for detection of air targets (where resolutioin is not crucial) is described. More recent work includes [4]–[6]. In [4] USRP B210 is mounted on an Unmanned Aerial Vehicle (UAV) to serve as GPR. The B210 bandwidth is insufficient to obtain an image of the subsurface, and, thus, just the object detection is described, with peaks marking the locations of the underground objects. In [5] the USRP X310 is used for detection of air targets. Note that for air targets radar resolution is not crucial, so the instantaneous SDR bandwidth of 160 MHz suffices. In [6] NI-USRP 2920 and USRP RIO are used for ice detection at large depths, so the radar resolution isn't critical as well. In [7] the SF approach, where the stepped frequencies span the available SDR frequency range, is implemented using an inexpensive SDR, the HackRF SDR. Our work closely parallels [8], based on USRP X310, where a number of subbands, each spanning the instantaneous RF bandwidth, are used to increase the radar bandwidth, and, thus, its resolution. Phase discontinuities that appear when the frequency changes are addressed in both [7] and [8], although in different ways.

Our goal is to design an SF Ground Penetrating Radar using an SDR device. As the SDR device here we use the Universal Software Radio Peripheral (USRP) X310, manufactured by the National Instruments Ettus company. In the SF approach [9], [10] pulses of different frequencies (the frequencies of the consecutive pulses differ by an equal amount - the frequency step) are sent, and the returns are measured, in order to

determine the amplitude and phase change at each frequency, i.e. the channel frequency response. The Inverse Discrete Fourier Transform (IDFT) then gives the range profile. It is well known that the radar range resolution depends on the bandwidth used. Since we are interested in designing a GPR radar for land mine detection (including both anti-tank and anti-personnel mines), the resolution has to be of the order of centimeters, and thus, the required bandwidth has to be several GHz. The main drawback of the SDR radar is the relatively narrow instantaneous RF bandwidth, which demands more complex approaches in order to obtain the needed radar range resolution. The SF radar is a solution to this problem, since the frequency can be hopped throughout the whole frequency range of the device. However, in SF radar, there is a significant time interval required for the transmitter and receiver to settle to the new frequency. To reduce the total time for generation of the radar signal, in [11] we developed a variation of the SF radar, where the baseband signal is also a collection of pulses with increasing frequencies (stepped frequency signal) which spans the entire instantaneous RF bandwidth after up-conversion to RF, i.e. the subband. The RF frequency is changed using the SF approach, to obtain all the subbands that cover the total used radar frequency bandwidth. To implement the radar, here we use a combination of the SF and OFDM principles. The baseband signal is an OFDM signal, transmitted in different subbands by modulating different RF frequencies, using the SF approach. The DFT coefficients evaluated from the received signals of all different subbands are concatenated in a single long vector. Then, taking an IDFT, the range profile is obtained. In order to improve the receive SNR, each burst (signal up-converted to a given RF frequency) is obtained by repeating a number of times the baseband OFDM signal.

The second drawback typical for SDR devices, is the random phase change that occurs whenever the RF frequency changes. The phase changes between the adjacent subbands are relatively easy to compensate, because they are manifested as first order discontinuities. However this can be only achieved with precise frequency, phase and time synchronization [12]. The novelty of this work compared to [11] consists in using a different baseband signal, i.e. OFDM, which provides faster signalling, uses a continuous time baseband signal, and simplifies the detection process. The signal repetition within each burst significantly improves SNR, providing improved system robustness. The difference from [8] is in using the standard OFDM representation of the baseband signal and in the approach to the compensation of phase discontinuties between the adjacent subbands. To eliminate phase discontinuities, after applying the techniques described in [12], the magnitude of the phase jump between the adjacent subbands is first determined, and, then, its compensation is performed. Field experiments were performed outdoors, with a school version of an anti-tank mine buried in the ground as the target. The obtained results show accurate detection and localization of the buried target. Additionally, we use the Singular Value Decomposition (SVD) approach for the direct and the ground reflected wave removal [13], [14].

This paper is organized as follows. In Section II, USRP hardware and GNU Radio software radar implementation are described. In Section III, the signal processing procedures of the transmit and receive signals are presented. In Section IV the results of the field experiments using the proposed radar implementation are shown.

II. Description of Hardware and Software

The SDR offers flexibility in the design and implementation of various radio systems. One of the popular tools for implementing SDR is the free and open source software radio development environment called GNU Radio, designed to operate on PC compatible hardware running (primarily) Linux. The SDR hardware used was NI Ettus X310 USRP with Xilinx Kintex-7 FPGA and UBX-160 RF daughtherboard with frequency range 10 MHz - 6 GHz and 160 MHz instantaneous bandwidth. The USRP was connected to a desktop PC via dual 10 Gigabit Ethernet card and PCI Express interface which enables sampling rates up to 200 MSamples/s (Full Duplex). We used GNU Radio version 3.14.1 with GRC (GNU Radio Companion) and UHD (USRP Hardware Driver) API version 3.14.0. As transmit and receive antennas we used a pair of AH Systems SAS-571 double ridge guide horn antennas that operate in the range from 700 MHz to 18 GHz with 1.4-15 dBi gain. To operate the SDR radar, a suitable flowgraph model was created in GRC, shown in Fig. 1.

Besides the standard blocks offered by the GNU Radio software, we created two additional Out-Of-Tree (OOT) blocks. The first OOT block serves for generating the desired baseband OFDM signal. The theoretical analysis is described in Section III, where we talk about the signal processing aspects. This block is represented by the ”Vector Source” block including an imported funcionality, represented by the block Import:freq16384. The ”Vector Source” block is connected with the ”Vector to Stream” block, which is then connected to ”Float to Complex” block so that the desired stream is created. This stream is then sent to the ”UHD USRP Sink” block which is the transmitter block. The receiver ”UHD USRP Source” block is connected with the ”File Sink” block which stores the received signal for further processing, described in Section III. The second OOT block (”Sweep”) is a control block that sets the desired carrier frequency for both the transmitter and receiver using PMTs (Polymorphic Types) through the flow graph. In this block the initial frequency and the frequency step are defined and controlled by the ”Message Strobe” block. The ”Message Strobe” defines the time instant when certain carrier frequency is set for both the transmitter and receiver. After that, a time period is added to enable that a new carrier frequency is set for both the transmitter and receiver. With extensive testing of this model it was concluded that the carrier frequency of transmitter and receiver can be changed every 10 ms without introducing errors.

The created flowgraph, with additional modification regarding the synchronization and controlling the transmitter and receiver, was used to perform the field measurements.

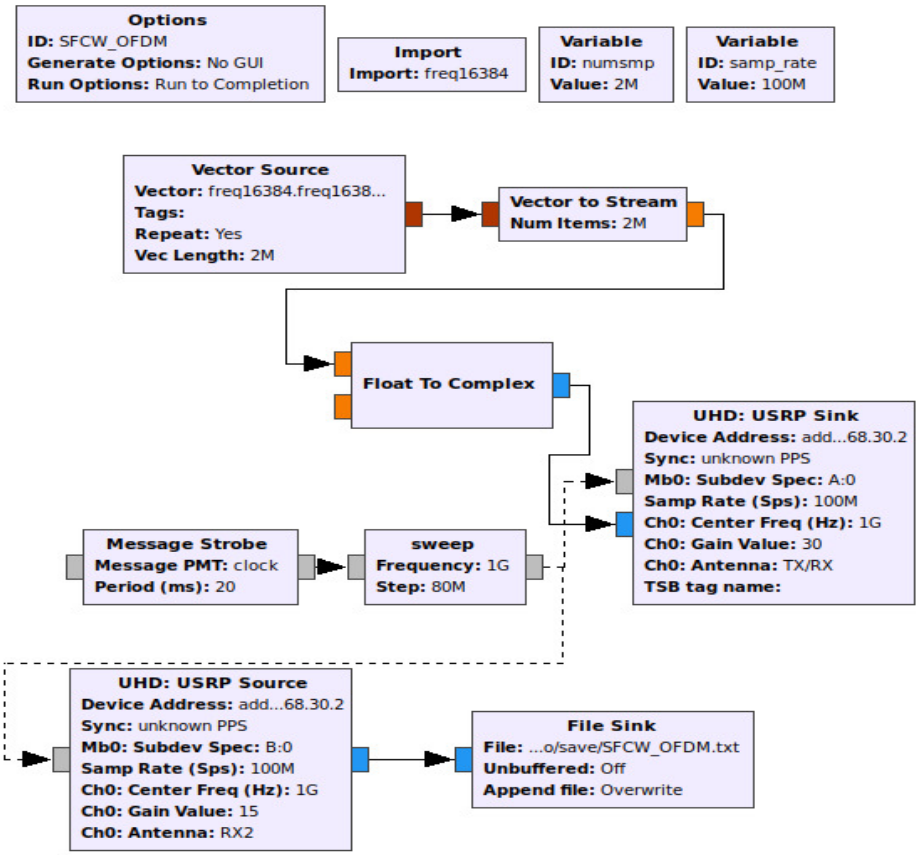


Fig. 1. GRC flowgraph of the implemented radar

The synchronization of the transmitter and the receiver was achieved with setting the exact start time on both transmitter and receiver. Additional delay was introduced in order to cope with the hardware/system delays and to provide the right amount of time for the USRP to properly set up.

III. Transmit and Receive Signal Processing

There are several types of scans produced by GPR. An A-scan is the range profile recorded at a given point above the ground surface (the position of the transmit and receive antennas) and is a one-dimensional signal in terms of range (or time) coordinate. A B-scan is a two dimensional image, obtained as a collection of concatenated A-scans, measured at equidistant locations on a straight line above the ground surface. In order to obtain the A-scans we use the following approach. Since the USRP has limited instantaneous bandwidth, we use a set of RF frequencies, modulated with the same baseband signal, thus forming subbands, and the total radar bandwidth is given by the collection of all these subbands. Thus, we can theoretically increase the total used bandwidth up to the intersection of the frequency range of the USRP and the frequency range of the antennas, i.e. in our case [700 MHz, 6 GHZ]. We use a baseband signal that is an OFDM signal. The OFDM signal is specified in the frequency domain as a collection of M equidistant non-zero frequencies, where the frequency spacing between two adjacent frequencies is Δf. To specify the resolution of the baseband OFDM signal we use a DFT of size $L > M$. Denote the L-dimensional vector that contains the amplitudes of the complex exponentials used in the system (while the phases are set to zero) with D. The positions of the desired M non-zero frequencies in the OFDM signal are specified with non-zero values (equal to one) in the vector D. The time domain baseband signal (base signal) is obtained using the Inverse Discrete Fourier Transform (IDFT) of D

$$d(n) = \frac{1}{L}\sum_{k=0}^{L-1} D(k) e^{\frac{j2\pi nk}{L}}, \; n = 0, \ldots, N-1 \qquad (1)$$

An example of the frequency domain OFDM signal is shown in Fig. 2, where the sampling frequency is 100 MSamples/s, the frequency step Δf=10 MHz and M=8. The frequency step is related to the radar unambiguous range as $R_u = v/(2\Delta f)$, which, for propagation in air (where $v = c$) and Δf=10 MHz

gives R_u=15 m. The time domain baseband signal from (1) is shown in Fig. 3.

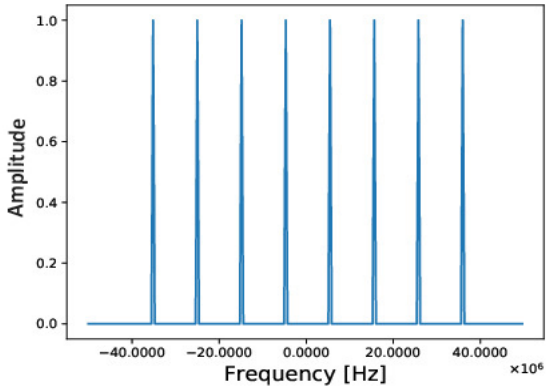


Fig. 2. Frequency domain baseband signal

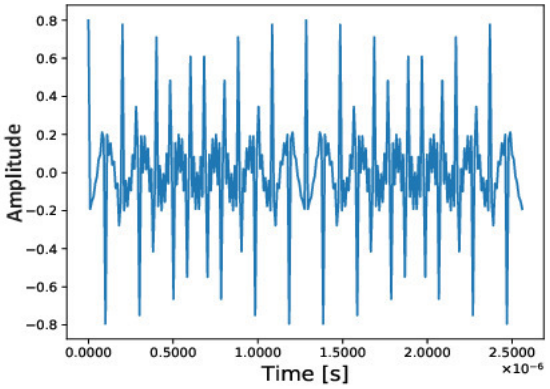


Fig. 3. Time domain baseband signal

The OFDM signal is then up-converted to a given RF frequency to create a subband. The RF frequencies are chosen in such a manner that the frequency spacing is constant throughout the entire used bandwidth, i.e. the frequency spacing between the highest non-zero frequency of a subband and the lowest non-zero frequency of the adjacent higher subband is also Δf. Every base signal is repeated K times before hopping to the next RF frequency. We call the obtained signal a burst. We do not use a guard interval (cyclic prefix) between different base signals within the burst, thus enabling faster transmission. Cyclic prefix is not needed, since the system is used for detecting buried targets that are close to the surface (AT mines buried in the ground), and, thus, the total time lag between the transmit and receive signals is less than one sample. By sending N bursts on N different RF frequencies, we equivalently send MN different non-zero frequencies (the repetitions excluded). By using the OFDM signal instead of the SF baseband signal as in [11], we speed up the transmission time and simplify the detection process. Also, since the OFDM signal is a continuous signal, this facilitates synchronization and improves the SNR. Moreover, the signal repetition provides significantly improved receive SNR, since the DFT coefficients obtained by taking the DFT of the received signal responses from the transmitted repeated signals within the burst can be averaged. The received time domain signal is first divided into chunks of length L, then converted to frequency domain using DFT of size L on each chunk, and identyfing the M non-zero coefficients. The signal is processed on a burst by burst basis, and the DFT coefficients from the K repetitions per burst are averaged. This is repeated for all N bursts. Thus, a long vector of $P = MN$ complex DFT coefficients is obtained, but phase discontinuities between adjacent subbands also appear. The phase response is shown in Fig. 4. To obtain accurate A-scans from the measurements, performing phase discontinuity compensation is crucial. To eliminate phase discontinuities after applying the techniques described in [12], the magnitude of the phase jump between the adjacent subbands is first determined, and, then, its compensation is performed. The phase response obtained after removing the phase discontinuities is shown in Fig. 5.

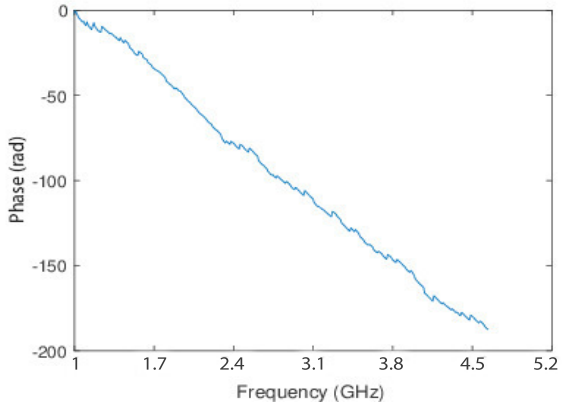


Fig. 4. Phase response before phase discontinuity compensation

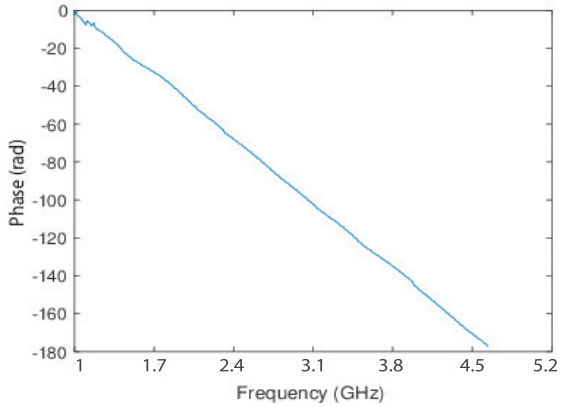


Fig. 5. Phase response after phase discontinuity compensation

IDFT is applied to the complex vector $\mathbf{X}_q$, made of elements $X_q(k) = A_q(k)e^{j\phi_q(k)}$ for $k = 0, \ldots, P-1$, where $q = 0, \ldots, Q-1$ is the index of the A-scan, and Q A-scans are

used to obtain the B-scan. The q-th A-scan ($q = 0, \ldots, Q-1$) is evaluated as follows:

$$a_q(n) = \frac{1}{P}\sum_{k=0}^{P-1} X_q(k) e^{\frac{j2\pi nk}{P}} = \frac{1}{P}\sum_{i=0}^{P-1} A_q(k) e^{j[\frac{2\pi nk}{P}+\phi_q(k)]} \tag{2}$$

for $n = 0, \ldots, P-1$, and forms a column vector $\mathbf{a}_q$. The B-scan is then the matrix:

$$A = [|\mathbf{a}_0|, \ldots, |\mathbf{a}_{Q-1}|] \tag{3}$$

However, the post-processing does not end with the phase doscontinuity compensation. The direct wave between the transmit and receive antennas and the ground reflected wave can have high energies, and these signals often mask the targets. They are parts of the clutter. In order to reduce the effects of the clutter, we propose using the Singular Value Decomposition (SVD), which we now briefly describe. The SVD of an $P \times Q$ image A (with $P > Q$) is given by

$$A = U\Sigma V^T = \sum_{i=0}^{Q-1} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{4}$$

where $U = [\mathbf{u}_0, \ldots \mathbf{u}_{P-1}]$ and $V = [\mathbf{v}_0, \ldots \mathbf{v}_{Q-1}]$ are the matrices of orthonormal basis vectors, i.e. the eigenvectors of AA^T and A^TA respectively (^{T} denotes matrix transpose). Σ is a $P \times Q$ matrix, with only nonzero elements $\{\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_Q\}$ obtained for equal row and column indices, which are called singular values of A. Since the singular values represent energy coefficients from the B-scan, in the image expansion in terms of the singular vectors, we exclude the term corresponding to the largest singular value in order to reduce the effects of the direct and the ground reflected waves.

IV. Results of the Experiments

The experiments were conducted outdoors, using the school version of the anti-tank (AT) mine TMM-1 (VTMM-1) as a buried object, shown in Fig. 6. The AT mine diameter is 30 cm, with height of 9 cm and weight of 8.6 kg, from which 5.6 kg of explosive content (TNT). The object was buried approximately 10 cm bellow the ground surface. The survey line was 190 cm long, and the antennas were positioned 50 cm above the ground. The experiment setup is shown in Fig. 7.

Fig. 6. The AT mine used as target

Fig. 7. The setup of the experiments

Two experiments were performed with different RF bandwidth and number of A-scans. The first experiment was performed using N=45 bursts, giving an effective RF bandwidth of B=3600 MHz. In this experiment Q=26 A-scans were recorded along the survey line, and the spatial step Δx between the A-scans was approximately 7 cm. The second experiment was performed using N=20 bursts, giving an effective RF bandwidth of B=1600 MHz. In the second experiment Q=15 A-scans were recorded, and the spatial step Δx between the A-scans was approximately 12.5 cm. The parameters used in both experiments are shown in TABLE I.

TABLE I – Parameters used in the experiments

Experiment	Bursts (N)	A-scans (Q)	B (MHz)
1	45	26	3600
2	20	15	1600

The DFT size of the OFDM signal was set to L=256 bins, out of which M=8 were non-zero frequencies. In both experiments the burst was created using K=64 repeated base signals. Both the range and spatial resolutions in the first experiment are superior to those used in the second experiment. This is illustrated in Figures 8 and 9.

The detected object, i.e. the AT mine (marked with yellow) is more blurred in the second experiment, because of the lower resolution. The lower resolution of the second experiment is also noticeable when we compare the B-scans obtained using the post processing SVD procedure for clutter removal, shown in Figures 10 and 11.

Despite the different resolution in the two experiments, the object is clearly visible in both of them. As mentioned in the previous sections, phase discontinuity compensation procedure should be applied on every A-scan of the B-scan, in order to obtain viable information. The A-scan corresponding to the phase response from Fig. 4 is shown in Fig. 12, and the A-scan after the phase discontinuity compensation (shown in Fig. 5) is shown in Fig. 13. It can be seen that the phase is non-linear in both cases, which is a consequence of the presence of multiple targets (ground surface and AT mine at the least). However, the abrupt changes in the phase response in Fig. 4

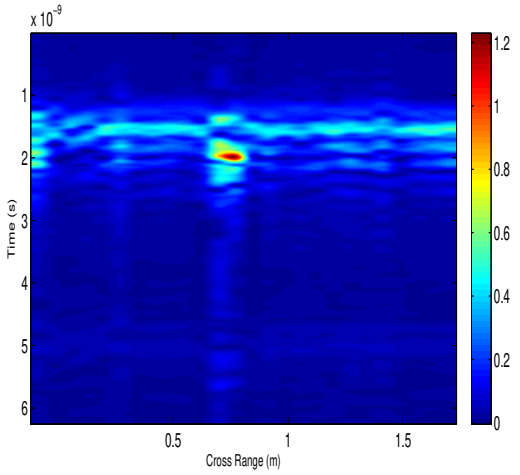


Fig. 8. B-scan first experiment

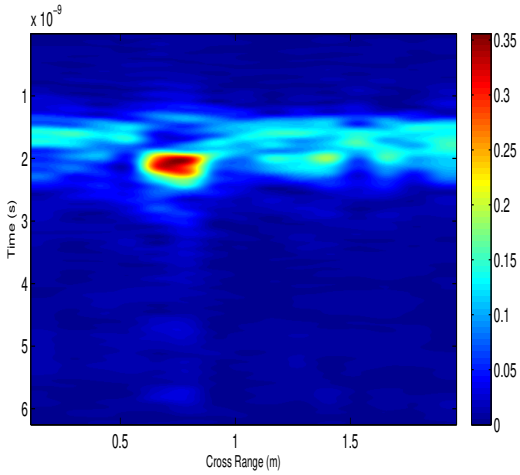


Fig. 9. B-scan second experiment

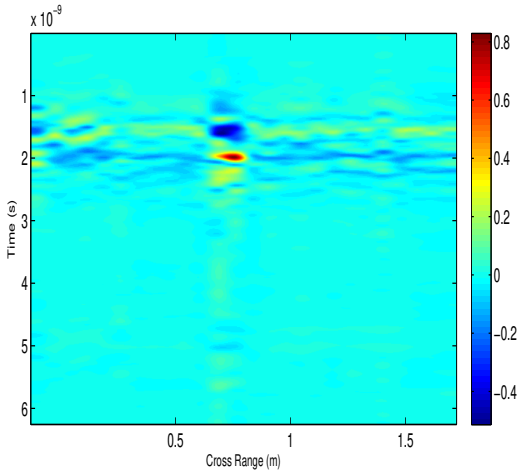


Fig. 10. B-scan with applied SVD first experiment

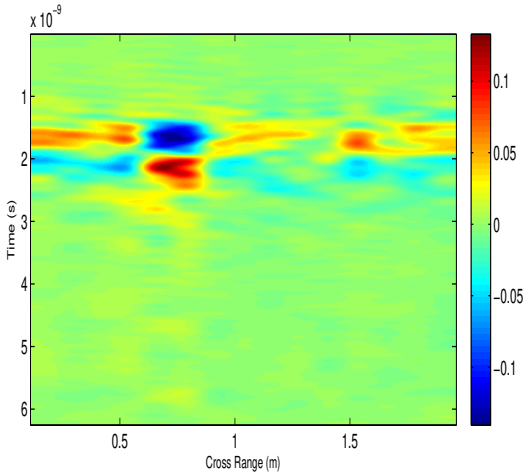


Fig. 11. B-scan with applied SVD second experiment

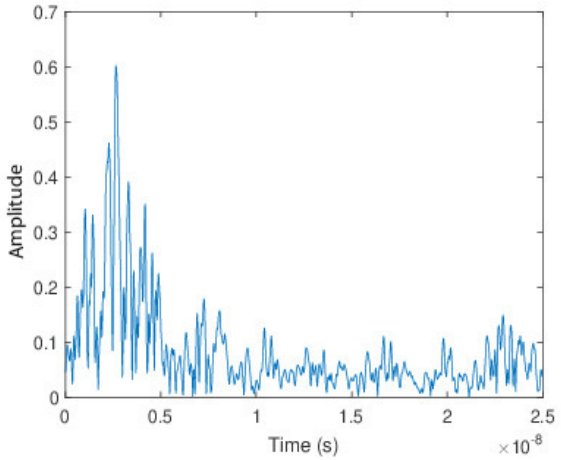


Fig. 12. A-scan before phase discontinuity compensation

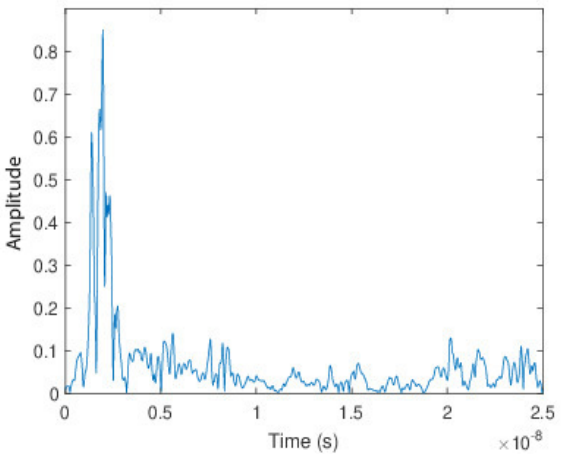


Fig. 13. A-scan after phase discontinuity compensation

cause imprecise range estimation and target masking, visible in Fig. 12, while the improved range profile is evident in Fig. 13. Imprecise range estimation in the A-scans causes B-scan image distortion, making the target detection impossible.

V. Conclusion

We describe an SF GPR implementation using USRP X310 and GNU Radio, that results in improved performance compared to our previous work [11]. The proposed system is specifically designed for practical implementation and addresses the SDR hardware imperfections and limitations. Using OFDM as a baseband signal speeds up the transmission, and improves and simplifies the detection process. The baseband signal repetition within each burst before upconverting to RF frequency significantly improves the receive SNR. We also propose a method for successful phase discontinuity compensation, that is crucial for proper radar operation. The B-scans obtained from the experimental data show excellent buried target detection capability, giving an accurate target location.

Acknowledgement

The authors would like to thank the NATO SPS Programme for the support of the presented work, provided by the project grant NATO SPS 985208.

References

[1] D. J. Daniels, Ground Penetrating Radar, 2nd Edition, Institution of Electrical Engineers, London, United Kingdom, 2004.

[2] J. C. Ralston, C. O. Hargrave, "Software defined radar: An open source platform for prototype GPR development," 14th International Conference on Ground Penetrating Radar (GPR), 2012.

[3] S. Costanzo, F. Spadafora, G. Di Massa, A. Borigia, A. Costanzo, G. Aloi, P. Pace, V. Loscr'i, H. O. Moreno, "Potentialities of USRP-based software defined radar systems," Progress In Electromagnetics Research B, Vol. 53, pp. 417-435, January 2013.

[4] M. P. Cerquera, J. D. Colorado, I. Mondragón, "UAV for Landmine Detection Using SDR-Based GPR Technology," Chapter from the book Robots Operating in Hazardous Environments, Edited by Hüseyin Canbolat, IntechOpen, 2017.

[5] K. Ranney, K. Gallagher, D. Galanos, A. Hedden, R. Cutitta, S. Freeman, C. Dietlein, B. Kirk, R. Narayanan, "Software-defined radar: recent experiments and results," Proceedings of SPIE - The International Society for Optical Engineering, ISSN: 1996-756X, Vol. 10633, pp. E1-E7, 2018.

[6] P. Liu, J. Mendoza, H. Hu, P. G. Burkett, J. V. Urbina, S. Anandakrishnan, S. G. Bilén, "Software-defined radar systems for polar ice-sheet research," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 12, No. 3, pp. 803–820, March 2019.

[7] J. Jendo, M. Pasternak, "Ground penetrating radar prototype based on a low-cost software defined radio platform," Przeglad Elektrotechniczny, Vol. R.95, No. 9, pp. 36-39, 2019.

[8] S. C. Carey, W. R. Scott, "Software defined radio for stepped-frequency, ground-penetrating radar," 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, pp. 4825-4828, 2017.

[9] D. R. Wehner, High-Resolution Radar, Second Edition, Artech House, 1995.

[10] C. Nguyen, J. Park, Stepped-Frequency Radar Sensors: Theory, Analysis and Design, Springer, 2016.

[11] V. Kafedziski, S. Pecov, "Implementation of a High Resolution Stepped Frequency Radar on a USRP," 13th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS), 2017.

[12] https://files.ettus.com/manual/page_sync.html

[13] B. Cagnoli, T. J. Ulrych, "Singular Value Decomposition and Wavy Reflections in Ground-penetrating Radar Images of Base Surge Deposits", Journal of Applied Geophysics, Vol. 48, pp. 175-182, 2001.

[14] V. Kafedziski, S. Pecov, D. Tanevski, "Target detection in SFCW ground penetrating radar with C3 algorithm and Hough transform based on gprMax simulation and experimental data," 25th International Conference on Systems, Signals and Image Processing (IWSSIP), 2018.